

**FLEXIBLE FORMANT SYNTHESIZER:  
A TOOL FOR IMPROVING SPEECH PRODUCTION QUALITY**

**BY  
AJIT LALWANI**

**A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL  
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT  
OF THE REQUIREMENTS FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY**

**UNIVERSITY OF FLORIDA**

**1992**

**UNIVERSITY OF FLORIDA LIBRARIES**

Copyright 1992

by

Ajit L. Lalwani

To

Mariko and my parents,  
who offered me love, inspiration and  
unfailing support.

## ACKNOWLEDGEMENTS

The author wishes to express his appreciation to his guru Dr. D. G. Childers, for his directions, financial and moral support, guidance and patience throughout the course of his graduate studies. His influence on the author may last for years to come. The author is grateful to Dr. J. R. Smith, Dr. J. C. Principe, Dr. A. A. Arroyo, and Dr. H. B. Rothman for serving on his supervisory committee and for the invaluable guidance provided to him. The author wishes to thank all the graduate students from the Mind-Machine Interaction Research Center for their help and patience.

The author is most grateful to his girlfriend, Mariko Sakai, for her moral support, encouragement, patience and help which cannot be fully expressed in words. The author is also grateful to his friends which include the Navathe, Marathe, Purchase, Gokhale families and Kedar Belhe for their constant encouragement and help. Last but not the least, the author is most grateful to his parents and his family members for their hard work, love, sacrifice and support, which has been the motivating force behind the author's accomplishments, life and work.



## TABLE OF CONTENTS

|                                                                       | <u>Page</u> |
|-----------------------------------------------------------------------|-------------|
| ACKNOWLEDGMENTS .....                                                 | iv          |
| ABSTRACT .....                                                        | ix          |
| <br>CHAPTERS                                                          |             |
| 1 INTRODUCTION .....                                                  | 1           |
| 1.1 Speech Synthesis Systems .....                                    | 1           |
| 1.2 Speech Synthesizers .....                                         | 2           |
| 1.2.1 Articulatory Synthesizers .....                                 | 4           |
| 1.2.2 LPC Synthesizer .....                                           | 5           |
| 1.2.3 Formant Synthesizers .....                                      | 6           |
| 1.3 Speech Quality .....                                              | 8           |
| 1.3.1 Intelligibility, Naturalness and Quality .....                  | 9           |
| 1.3.2 Quality of Synthetic Speech .....                               | 10          |
| 1.3.3 Assessing the Quality of Synthesized Speech .....               | 11          |
| 1.4 Voice Quality .....                                               | 14          |
| 1.5 Research Goals and Plan .....                                     | 16          |
| 1.6 Description of the Following Chapters .....                       | 19          |
| 2 A FLEXIBLE FORMANT SYNTHESIZER .....                                | 22          |
| 2.1 Introduction .....                                                | 22          |
| 2.2 Classification of Speech Sounds .....                             | 22          |
| 2.3 Flexible Formant Synthesizer .....                                | 25          |
| 2.3.1 Need for a Flexible Formant Synthesizer .....                   | 25          |
| 2.3.2 Development of a Flexible Formant Synthesizer .....             | 27          |
| 2.3.3 Synthesizer Block Diagram .....                                 | 29          |
| 2.3.4 List of Synthesizer's Parameters List .....                     | 29          |
| 2.3.5 Synthesizer Software/Hardware and Flowchart .....               | 35          |
| 2.3.6 Assessing the Performance of Flexible Formant Synthesizer ..... | 35          |
| 2.4 Basic Features of Flexible Formant Synthesizer .....              | 36          |
| 2.4.1 Digital Implementation in Time Domain .....                     | 36          |
| 2.4.2 Control Parameters of the Flexible Formant Synthesizer ..       | 38          |

|       |                                                                       |     |
|-------|-----------------------------------------------------------------------|-----|
| 2.4.3 | Glottal Source Models                                                 | 42  |
| 2.4.4 | Noise Source                                                          | 43  |
| 2.4.5 | Modulation of Noise Source                                            | 45  |
| 2.4.6 | Frequency-Shaping of Noise Source                                     | 47  |
| 2.4.7 | First Order Systems                                                   | 47  |
| 2.4.8 | Filter Banks                                                          | 51  |
| 2.5   | Radiation Load                                                        | 79  |
| 2.6   | Synthesis Strategies                                                  | 81  |
| 2.6.1 | Voiced Sounds                                                         | 81  |
| 2.6.2 | Examples for Specifying Parameters to Flexible Formant Synthesizer    | 87  |
| 2.6.3 | Unvoiced Sounds                                                       | 87  |
| 2.6.4 | Mixed Excitation Sounds                                               | 91  |
| 2.7   | Summary                                                               | 98  |
| 3     | ADVANCED FEATURES OF THE FLEXIBLE FORMANT SYNTHESIZER                 | 99  |
| 3.1   | Introduction                                                          | 99  |
| 3.2   | Flexible Formant Synthesizer                                          | 99  |
| 3.2.1 | Formant Tracking and Formant Synthesis                                | 99  |
| 3.2.2 | Problems with Flexible Configuration of Filter Banks                  | 101 |
| 3.2.3 | Solutions to the Problems with Flexible Configuration of Filter Banks | 102 |
| 3.2.4 | Configuration of the Filter Banks in the Flexible Formant Synthesizer | 113 |
| 3.2.5 | Examples                                                              | 116 |
| 3.3   | Sampling Rate of Synthesizer Speech                                   | 119 |
| 3.4   | Time and Frequency Domain Scaling                                     | 121 |
| 3.4.1 | Frequency Compression/Expansion                                       | 124 |
| 3.4.2 | Time Scaling/Variable Speaking Rate                                   | 134 |
| 3.5   | Source-Tract Interaction                                              | 132 |
| 3.5.1 | Glottal Pulse Skewing                                                 | 133 |
| 3.5.2 | Truncation of First Formant                                           | 133 |
| 3.5.3 | Tracheal Poles and Zeros                                              | 133 |
| 3.5.4 | Simulation of Source-Tract Interaction with the Formant Synthesizer   | 134 |
| 3.6   | Summary                                                               | 143 |
| 4     | GLOTTAL SOURCE MODEL                                                  | 144 |
| 4.1   | Introduction                                                          | 144 |
| 4.2   | Modelling Vocal Characteristics and Formant Synthesis                 | 144 |
| 4.3   | Glottal Flow Characteristics                                          | 145 |
| 4.3.1 | Time Domain Glottal Factors                                           | 146 |
| 4.3.2 | Frequency Domain Glottal Factors                                      | 147 |

|       |                                                                                 |     |
|-------|---------------------------------------------------------------------------------|-----|
| 4.4   | Vocal Characteristics and Glottal Factors .....                                 | 147 |
| 4.5   | Glottal Source Model .....                                                      | 150 |
| 4.5.1 | Voicing Source Model .....                                                      | 155 |
| 4.5.2 | Aspiration Noise Source Model .....                                             | 159 |
| 4.5.3 | Pitch Perturbation Source Model .....                                           | 169 |
| 4.5.4 | Amplitude Perturbation Source Model .....                                       | 181 |
| 4.6   | Summary .....                                                                   | 184 |
| 5     | GLOTTAL SOURCE MODEL AND GLOTTAL FACTORS .....                                  | 185 |
| 5.1   | Introduction .....                                                              | 185 |
| 5.2   | Controlling Time Domain Glottal Factors .....                                   | 185 |
| 5.2.1 | Controlling the Pitch Period ( $t_0$ ) .....                                    | 185 |
| 5.2.2 | Controlling the Glottal Source Pulse Skewness<br>(Speed Quotient, SQ) .....     | 187 |
| 5.2.3 | Controlling the Glottal Source Pulse Width<br>(Open Quotient, OQ) .....         | 187 |
| 5.2.4 | Controlling the Abruptness of Closure of the<br>Glottal Source Pulse (AC) ..... | 189 |
| 5.2.5 | Controlling the Aspiration Noise<br>(Signal to Noise Ratio, SNR) .....          | 193 |
| 5.2.6 | Controlling "Jitter" Using the New Glottal Source Model .                       | 193 |
| 5.2.7 | Controlling "Shimmer" Using the<br>New Glottal Source Model .....               | 196 |
| 5.3   | Controlling Frequency Domain Glottal Factors .....                              | 196 |
| 5.3.1 | Controlling the Fundamental Frequency .....                                     | 197 |
| 5.3.2 | Controlling the Spectral Tilt .....                                             | 197 |
| 5.3.3 | Controlling the Harmonic Richness Factor (HRF) .....                            | 200 |
| 5.3.4 | Controlling the Harmonic to Noise Ratio (HNR) .....                             | 203 |
| 5.4   | Frequency Domain Approach to Modeling Vocal Disorders ...                       | 208 |
| 5.5   | Summary .....                                                                   | 218 |
| 6     | EXPERIMENTS AND PERCEPTUAL EVALUATIONS .....                                    | 219 |
| 6.1   | Introduction .....                                                              | 219 |
| 6.2   | Experiments .....                                                               | 219 |
| 6.2.1 | Objectives .....                                                                | 220 |
| 6.2.2 | Method .....                                                                    | 220 |
| 6.2.2 | Results .....                                                                   | 223 |
| 6.2.2 | Conclusions .....                                                               | 225 |
| 7     | DISCUSSION .....                                                                | 226 |
| 7.1   | Summary .....                                                                   | 226 |
| 7.1.1 | The Flexible Formant Synthesizer .....                                          | 226 |
| 7.1.2 | New Glottal Source Model .....                                                  | 229 |
| 7.2   | Future Work .....                                                               | 231 |

|                                                                 |     |
|-----------------------------------------------------------------|-----|
| 7.2.1 Improvements in the Flexible Formant Synthesizer .....    | 232 |
| 7.2.2 Improvements in the New Glottal Source Model .....        | 233 |
| <b>APPENDICES</b>                                               |     |
| A FORMANT SYNTHESIZER ARCHITECTURE .....                        | 235 |
| Literature Survey of Formant Synthesizers .....                 | 235 |
| Limitations of the Current Formant Synthesizers .....           | 244 |
| B EXCITATION SOURCE: GLOTTAL AND NOISE .....                    | 248 |
| Glottal Source Models .....                                     | 248 |
| Noise Sources .....                                             | 253 |
| C SOFTWARE AND FLOW CHART .....                                 | 270 |
| Software/Hardware Description .....                             | 270 |
| Flowchart for Formant Synthesizer .....                         | 273 |
| D CONTROLLING JITTER IN SYNTHETIC SPEECH .....                  | 279 |
| Controlling the JF and FPQ Measures .....                       | 282 |
| Controlling the DJ Measure .....                                | 288 |
| Effect of Varying the DJ Measure on the JF and FPQ Measures ... | 290 |
| E HARMONIC TO NOISE RATIO .....                                 | 298 |
| REFERENCES .....                                                | 311 |
| BIOGRAPHICAL SKETCH .....                                       | 318 |

Abstract of Dissertation Presented to the Graduate School  
of the University of Florida in Partial Fulfillment of the Requirements for the  
Degree of Doctor of Philosophy

**FLEXIBLE FORMANT SYNTHESIZER:  
A TOOL FOR IMPROVING SPEECH PRODUCTION QUALITY**

By

Ajit Lalwani

May, 1992

Chairman: Dr. D.G. Childers  
Major Department: Electrical Engineering

A flexible formant synthesizer is required to develop commercial applications and conduct research with synthetic speech. Current implementations of formant synthesizers have limited parameter sets, inflexible synthesis algorithms and rigid architectures. Our implementation of the formant synthesizer is an outgrowth of Klatt's cascade/parallel formant synthesizer.

Additional parameters were incorporated to control the synthesis algorithms, configure the synthesizer architecture and specify acoustic parameters. Several glottal source and noise source models were provided for source excitation. Both the cascade and parallel filter banks were configured by simple parameter specifications. Simple algorithms were used to 1) specify a variable number of filters in both the filter banks at the start-up and during the synthesis, 2) simulate the cascade filter bank by the parallel filter bank, 3) create "zeros" in the magnitude frequency response of the parallel filter bank, 4) detect and remove "clicks" and "pops" which commonly occur in synthetic speech, 5) perform time and frequency scaling of an utterance and 6) simulate source-tract interaction phenomenon. The new/modified features of the

flexible formant synthesizer were illustrated with examples of synthesis of sustained vowels and sentences.

The flexible formant synthesizer was utilized for modeling/synthesizing various voice types. The advantage of this approach is that the acoustic characteristics significant for modeling/synthesizing various voice types can be properly controlled. This study focused on modeling/synthesis of creaky, breathy, modal, rough and hoarse voice types. A new glottal source model was developed and incorporated in the flexible formant synthesizer. This model has 1) a voicing source, 2) an aspiration noise source, 3) a pitch perturbation source and 4) an amplitude perturbation source. The time domain glottal factors, such as pitch-period, glottal pulse width, glottal pulse skewness, abruptness of closure of the glottal pulse, aspiration noise, jitter and shimmer, and the frequency domain glottal factors, such as spectral tilt, Harmonic Richness Factor and Harmonic to Noise Ratio were controlled by the glottal source model's parameters. Relationships between glottal factors and the glottal source model's parameters were illustrated with graphs and analytical expressions.

Speech tokens (sustained vowels and sentences) were synthesized, and the quality of these speech tokens were evaluated by informal listening tests. The preliminary results indicate that the glottal source model and the flexible formant synthesizer have the potential for improving the quality of synthetic speech with various voice types.

## CHAPTER 1 INTRODUCTION

Speech is the primary mode of human communication. Historically, the technological progress in speech communication has been focused mainly on the extension of the range and reliability of transmission of the speech signal. Currently, machines are being employed to synthesize speech instead of simply transmitting the speech signal. With the advent of high speed digital computers high-quality real-time speech synthesis is possible. Speech synthesizers are used in commercial applications such as telecommunication, voice-response-systems, aids for the blind and hearing-impaired, toys, etc. Several areas of speech research such as rule-based speech synthesis, speech analysis-synthesis, speech disorders, etc., employ speech synthesis to evaluate the performance of proposed systems and to obtain "feedback" for improvement. Other applications of synthetic speech are to fabricate speech stimuli for psychoacoustic experiments, and to study auditory and speech disorders in speech clinics.

### 1.1 Speech Synthesis Systems

Speech synthesis is the production of speech by a machine using algorithms, rules and acoustic parameters. The reproduction of speech from tape recorders, compact discs and other means is not considered as speech synthesis. Speech can be represented with waveforms of speech sounds or in terms of parameters of a speech production model. A typical speech synthesis system consists of a combination of a speech waveform parser or a parameter estimator and a speech synthesizer. A review of

several speech synthesis systems is given in Bristow (1984). The three most popular types of speech synthesis systems are as follows:

- 1) Direct-synthesis systems [Stella, 1985]: The portions of waveforms of smaller units of speech sounds, such as di-phones (two adjacent phonemes), demi-syllables (two adjacent syllables), words, etc., are obtained from the waveforms of larger units of speech, such as words or sentences. These waveforms are then concatenated together to synthesize several different speech waveforms.
- 2) Analysis-synthesis systems [Pinto et al., 1989]: The parameters of a speech production model are estimated by analyzing natural speech signals. These parameters are then used to replicate the original or to produce a modified speech signal using the speech production model.
- 3) Rule-based (text-to-speech) synthesis systems [Klatt, 1987]: The input text in the form of alphanumeric symbols is converted to its phonetic transcription (representation of text in terms of phonemes and their allophones (variation in sound of a phoneme)) by applying the phonetic and linguistic rules. The typical values of the parameters of a speech production model, for each phoneme and allophone, are obtained from parameter databases. These parameters are used to synthesize the speech signal from the speech production model.

The direct-synthesis systems are simple but the quality and intelligibility of the synthesized speech is tolerable only for limited applications, such as toys, inexpensive answering services, etc. The analysis-synthesis and the rule-based synthesis systems can produce intelligible and high-quality speech for a large number of applications.

## 1.2 Speech Synthesizers

In a speech synthesis system, a speech synthesizer generates the speech signal either from the estimated parameters of a speech production model or from



waveforms of the smaller units of speech sounds. Accordingly, the speech synthesizers are classified into two types as follows:

- 1) Direct synthesizers: This type of synthesizer generates speech by smooth concatenation of the stored waveforms of di-phones, demi-syllables, words, etc. These synthesizers are used in the direct-synthesis systems.
- 2) Parametric synthesizers: This type of synthesizer generates speech using a speech production model. These synthesizers are used in analysis-synthesis and rule-based synthesis systems.

The parameters and the architecture of the “parametric synthesizer” depends upon the speech production model it simulates. There have been two basic approaches used to model the human speech production system. Accordingly, the parametric speech synthesizers have been categorized into the following two groups:

- a) Direct approach: These speech synthesizers are based upon a speech production model that directly simulates the physiological and acoustical aspects of the human speech production and propagation systems. The articulatory synthesizers belong to this group.
- b) “Terminal-Analog” approach: The speech synthesizers are based upon a speech production model that attempts to simulate the acoustical aspects of human speech production and propagation systems from an input and output point of view. The synthesizer is treated as a “black box” that simulates the acoustical characteristics of speech. The synthesizer parameters serve as input to the “black-box” with the output being speech. The formant and LPC (linear predictive coding) synthesizers belong to this group.

Selection of the appropriate synthesizer for an application largely depends upon a “best fit” method in which the synthesizer’s advantages and disadvantages are weighed against each other with respect to the nature of the research and applications.

The following subsections briefly describe the parametric synthesizers, and their advantages and disadvantages.

### 1.2.1 Articulatory Synthesizer

Articulatory synthesizers are based upon a speech production model that simulates the source of speech sounds by the movements of the vocal folds and other articulators, and also simulates the aerodynamic propagation of speech sounds through the vocal and nasal tracts [Flanagan 1972]. The data for the vibratory patterns of the vocal folds and the shapes of the vocal and nasal tracts are supplied by the vocal fold and articulatory models, respectively. The vocal fold models attempt to compute the glottal flow from such parameters as sub-glottal pressure, time-varying glottal area and the time-varying shapes of the vocal and nasal tracts [Flanagan and Landgraf, 1968; Ishizaka and Flanagan, 1972]. Since the generation of the glottal flow (source) includes the time-varying nature of the shape of the vocal and nasal tracts (acoustic load), the excitation source and the vocal tract characteristics are not considered independent of each other. Articulatory models specify the cross-sectional area of the vocal and nasal tracts along the mid sagittal plane. These area functions are used to compute the transfer functions of the vocal and nasal tracts, which simulate the propagation of sound through the vocal and nasal tracts, respectively. The area functions of the vocal tract and the nasal tract are interpolated between the target values (typical area functions for sustained phonations) of adjacent phonemes in order to synthesize continuous speech.

The advantages of articulatory synthesizers are as follows:

- 1) The parameters of articulatory synthesizers are directly related to the articulatory mechanisms, making them a valuable tool for speech production and perception studies.

2) Articulatory synthesizers simulate the motion of the vocal folds and the vocal tract as one system. Therefore, source-tract interaction, which is considered an essential factor for synthesizing natural-sounding speech, can be properly modeled.

The disadvantages of articulatory synthesizers are as follows:

- 1) It is difficult to obtain the required area functions of the vocal tract and vocal folds from human subjects. The area-functions obtained by X-ray photographs provide only the target values of a few phonemes and do not indicate the variations in the area-function due to "adaptations" or "coarticulation."
- 2) Considerable calculations are required to solve the system equations (transfer function) of the articulatory model. Also, the numerical analysis methods used for solving the system equations have stability problems when the solutions contain both slow and fast varying components.

### 1.2.2 LPC Synthesizer

LPC synthesizers [Atal and Hanauer, 1971] are based upon the source-filter speech production model proposed by Fant (1960). In LPC synthesizers, each new speech sample is synthesized from a weighted linear combination of the past synthesized speech samples and from the current samples of the excitation source. The LPC synthesizers can be implemented with an excitation source and a time-varying "all-pole" filter bank. Traditionally, the excitation source used for LPC synthesizers has been an impulse train for voiced sounds and white-noise for unvoiced sounds. Recently, Atal and Remde (1982) have used a multi-pulse excitation technique, and Schroder and Atal (1985) have used stochastically coded waveforms as an excitation source to improve the quality of LPC synthesized speech. Childers and Wu (1990) have used various stylized pulses as an excitation source to demonstrate an improvement in the "quality" of the LPC synthesized speech.

The advantages of the LPC synthesizers are as follows:

- 1) Fewer parameters are required to control the synthesizer, and therefore, LPC synthesizers are useful for speech coding and telecommunication applications.
- 2) Fast algorithms are available for calculating LPC coefficients.
- 3) The synthesized speech is intelligible even at bit rates as low as 4.8 KB/s.

The disadvantages of LPC synthesizers are as follows:

- 1) The parameters have little or no relation to the anatomy and physiology of the human speech production system.
- 2) The source-tract interaction is not simulated in a direct manner.
- 3) The synthesized speech exhibits unnatural characteristics.
- 4) The "quality" of nasals and unvoiced sounds is poor.

### 1.2.3 Formant Synthesizer

Formant synthesizers are also based upon the source-filter speech production model [Fant, 1960]. The source models provide stylized volume-velocity waveforms to excite the filter banks. The filter banks consist of resonators and anti-resonators that model the time-varying frequency domain transmission characteristics of the vocal tract, i.e., the transfer function of the vocal tract relates the volume-velocity at the glottis to the volume-velocity at the lips. The filter bank(s) contribute to the envelope of the short-time magnitude frequency response of the vocal tract transfer function, while the nature of the harmonics of the source and its spectral tilt are controlled by the excitation waveform. The lip radiation (conversion of volume-velocity at the lips to speech pressure wave) is simulated by a simple filter.

The vocal tract transfer function can be reduced to either a cascade of complex pole-pair networks (resonators) or a parallel addition of complex pole-pair networks [Flanagan, 1957]. Although there is a controversy over which filter bank configuration accurately simulates the frequency domain transmission characteristics of the vocal

tract for voiced sounds, the cascade/parallel filter bank configuration proposed by Klatt (1980) is the most popular one.

The similarity between the LPC and the formant synthesizers is that the peaks in the magnitude frequency response of the filter banks in both synthesizers match the peaks in the spectra of the original speech sounds. The differences between the LPC and formant synthesizers are as follows:

- 1) A single filter bank in LPC synthesizers simulates the combined frequency domain characteristics of the source, the vocal tract and the lip radiation, whereas the filter bank(s) in the formant synthesizer simulate the frequency domain characteristics of the vocal tract alone.
- 2) The LPC synthesizer employs an IIR (Infinite Impulse Response) filter bank. The formant synthesizer employs filter bank(s) consisting of second order resonators and anti-resonators. Also, the procedures followed to obtain the filter coefficients are different for each synthesizer (although the resonator coefficients can be obtained from the LPC coefficients, and vice versa).

The advantages of the formant synthesizers are as follows:

- 1) The formant synthesizer is more closely related to the human speech production and sound propagation systems than the LPC synthesizer and is not as complex as the articulatory synthesizer. Thus, the formant synthesizer has a higher potential for producing high-quality synthesized speech than an LPC synthesizer. Although the articulatory synthesizer has a potential for synthesizing high-quality speech, the research needed for improving its quality is presently limited by the existing models and the number of computations required for extracting the system parameters from the speech signal.
- 2) The parameters of the formant synthesizer are closely related to the spectral and acoustical properties of the speech sounds. Therefore, the formant synthesizer can be

used for studying the inter-relationships between the speech production and perception processes.

- 3) Existing formant synthesizers produce high-quality speech, if the synthesizer parameters are carefully controlled.
- 4) Factors responsible for quality, such as formant locations, formant bandwidths and excitation waveforms, etc., can be independently controlled.
- 5) The effect of source-tract interaction can be simulated by modifying the shape of the glottal pulses generated by the glottal source model. Also, the formant frequencies and formant bandwidths can be modified for the duration corresponding to the open phase of vocal fold vibrations.

The disadvantages of the formant synthesizers are as follows:

- 1) Obtaining the formant tracks (formant frequencies, bandwidths and amplitudes) from the speech signal is very difficult due to artifacts generated by the methods employed.
- 2) The speech produced by formant synthesizers sounds too "smooth." The formants, anti-formants and the excitation source parameters are considered as piece-wise linear and are varied slowly during the synthesis of an utterance. The fast transitions in the speech sounds are either ignored or are smoothed by the formant tracking algorithms.

### 1.3 Speech Quality

With the increasing use of synthetic speech for communications and other commercial and research applications, the generation of natural sounding synthetic speech is becoming a requirement for a successful speech synthesizer in the market place. Increasing demands have been placed on the the finer attributes of synthetic speech, such as intelligibility, quality, recognizability and naturalness, for synthesizing

“human sounding” speech: As seen from the previous sections, a considerable amount of research effort has been expended to develop of speech synthesizers and to improve the quality of synthetic speech. However, the quality of the synthetic speech generated by most of the speech synthesizers has been rated “inferior” as compared to the “natural” speech. The synthetic speech has often been described as either being “buzzy,” “bassy,” “metallic,” or “monotonic.” As speech synthesis techniques continue to develop, people will become increasingly discontent with the “inferior” synthetic speech quality, and it is precisely this discontentment that has signaled the need for improvement in the quality of synthetic speech.

### 1.3.1 Intelligibility, Naturalness and Quality

Childers and Wu (1990) have noted that often speech researchers have used the terms “quality,” “intelligibility” and “naturalness” interchangeably. For our purpose, the specific usage of these terms in the speech literature is as follows:

#### 1.3.1.1 Intelligibility

The intelligibility of speech is related to the ability of a listener to correctly identify the units of speech stimuli such as phonemes, syllables, words or sentences, given that the language is known to the listener and that the syntax and semantics are correct.

#### 1.3.1.2 Naturalness

Naturalness is used synonymously for the impression “human-sounding.” This definition has highly subjective attributes since the listener’s impression is influenced by several factors such as clarity, speaker’s age, speaking rate, dialect, accent, background noise, etc.

### 1.3.1.3 Quality

This term has a different meaning in different contexts. To describe a normal speech dialogue or monologue, the terms “quality” and “naturalness” are used interchangeably to describe naturalness of human speech. The term “quality” is used by speech phoneticians to describe articulatory differences when comparing the vowels in different words. Speech pathologists may use “laryngeal quality” to describe various voice characteristics, such as hoarse, harsh, breathy, etc. A singer may use the term “quality” to express differences in vocal registers related to laryngeal vibratory characteristics, such as vocal fry, modal and falsetto.

### 1.3.2 Quality of Synthetic Speech

Rothauser et al. (1971) have described speech quality in terms of four factors: 1) loudness, 2) speaker recognizability, 3) intelligibility and 4) preference. Preference becomes a dominant factor of speech quality when the loudness is at a comfortable level, intelligibility is good and speaker recognizability is of no interest. They have suggested that the listener’s preference may be expressed as the proportion of the listening group that prefers the speech test signal to the speech reference signal as a source of information. The listeners should be capable of expressing their preference consistently. Childers and Wu (1990) have used this definition for assessing the quality of synthetic speech produced by LPC and formant synthesizers. The listeners were asked to consider quality as equivalent to naturalness during the preference tests.

Several aspects of this definition of speech quality should be considered when it is applied to define quality of synthetic speech. For synthesized speech the loudness can be adjusted by simple scaling of the sampled speech waveform or by adjustment of the volume control of the playback amplifier, and therefore, it is not an important factor to be considered during speech synthesis. When communicating by speech,



listeners may be interested in several factors and not just speaker recognizability. For example, in telecommunication, where synthesizers are used as vocoders, the listeners may be interested in such factors as the emotional state, the accent, etc. of the speaker. The speech synthesizers developed to date are incapable of reproducing most of these factors in synthetic speech. Finding the acoustic features significant for reproducing these factors in synthetic speech is a current research area. Both the intelligibility and naturalness of the synthesized speech tokens can be assessed by evaluating a listener's preference in the listening tests. While evaluating only the naturalness of synthetic speech, its intelligibility should be high. While evaluating only the intelligibility of synthesized speech, its naturalness should be comparable with that of the human speech.

We consider "high-quality" synthetic speech as both highly-intelligible and natural-sounding, and consider a "high-quality" speech synthesizer as one that is capable of producing highly-intelligible and natural-sounding speech.

### 1.3.3 Assessing the Quality of Synthesized Speech

In the literature on speech quality we find that several researchers have attempted to assess the "intelligibility," "naturalness" and "quality" of synthetic speech using both qualitative and quantitative methods. Childers and Wu (1990) have described various qualitative and quantitative methods used by several speech researchers. In this section we give a general description of qualitative and quantitative methods and discuss their pros and cons.

#### 1.3.3.1 Qualitative methods

This method involves listening tests in which the intelligibility and/or naturalness of synthesized speech is perceptually evaluated by a group of judges, i.e., by listeners

trained to attend to specific aspects of the speech signal. The objective of the listening tests is to compare the intelligibility and/or naturalness of the reference speech tokens with test speech tokens. The judges selected for the listening tests should have:

- 1) a mutual agreement on the definitions of various acoustic features in the speech signal and their perceptual correlations with speech quality,
- 2) training in estimating the values of various acoustic features in the speech tokens and a demonstrated ability to discriminate the quality of speech tokens based on the similarities or differences in the values of these parameters, and
- 3) ability to express their evaluations consistently.

Since all these criteria may not be met by a given group of judges, a listening test may be highly subjective in nature, i.e., biased by the judge's opinions.

The listening tests can be conducted in a "formal" or "informal" manner. Formal listening tests require a large set of phonetically balanced speech tokens, generated by systematic variations of the pre-defined, perceptually significant acoustic features. Human subjects or speech synthesizers are employed to generate the phonetically balanced set of speech tokens. The listening test consists of an audio presentation of the reference and test speech tokens to the experienced judges. The judges are asked to rate the quality (intelligibility and naturalness) of the speech tokens on a pre-defined scale. The informal listening tests may use only a few speech tokens. The judges may not be asked to rate the quality of the speech tokens on a scale; instead, they may be asked to give general remarks about the quality of the speech tokens.

#### 1.3.3.2 Quantitative methods

From a technical point of view one prefers to have an objective method for assessing the quality (intelligibility and naturalness) of synthesized speech because the results obtained by such methods are presumably reproducible. A distance metric or a distortion measure based on the acoustic features of the speech signal is the

foundation for these objective measures. Two speech tokens are compared using the same acoustic feature contours extracted from each speech token. A typical distance measure calculates the normalized sum of the squares of the differences in the corresponding values of the two contours. The overall distance between the reference and the test speech tokens is generally the sum or average of the weighted sum of the individual contour distance measures. The weight of each individual distance measure is chosen on the basis of its effectiveness for discriminating quality. By selecting a perceptually consistent set of acoustic features for measuring the overall distance, one hopes to achieve a high degree of correlation between the quantitative (objective) distance measures and listeners ratings of the same speech tokens.

The basic requirements for a distance measure are illustrated using three feature contours  $x$ ,  $y$  and  $z$ , as follows:

- 1)  $d(x,y) = d(y,x)$  symmetry
- 2)  $d(x,y) > 0$  for  $x \neq y$  positive definiteness
- 3)  $d(x,x) = 0$
- 4)  $d(x,z) < d(x,y) + d(y,z)$  triangular inequality
- 5)  $d(x,y)$  has a physically meaningful interpretation in the frequency domain.

A review of several distance and distortion measures is given in Eskenazi (1988). Some of the distance measures proposed in the literature are not symmetric and do not satisfy the triangular inequality, and thus, are not distance measures per se. Typical distance measures use the speech spectrum, cepstrum coefficients and log likelihood ratios of spectra, etc., as the acoustic features for comparing a test token to a reference token. The most frequently used distance measure is perhaps the Itakura-Saito distance measure [Markel and Gray, 1976].

The results of the quantitative measures attempt to agree with the results of the qualitative methods. The quantitative methods developed to date are incapable of capturing the sensitivity of listeners to the various acoustically significant features in

the speech signal, and therefore, do not generally agree with the listener's evaluations. This may be due to several factors. A distance measure may use inappropriate weights or an insufficient number of acoustic parameters. Another factor is the sensitivity of the listeners to changes in the values of acoustic features in some cases but not in others. For example, occasionally a decrease in the magnitude of an acoustic parameter by a few decibels may be perceived by the ear quite easily but not so at other times. This may be because of the masking of the perceptual significance of some acoustic parameters. Also, a distance measure calculates the normalized sum of the squares of the differences in the values of acoustic feature contours for all the frames of two speech tokens, and therefore, is not sensitive to localized differences in acoustics features. However, the human ear may be sensitive to localized differences in the contours of some acoustic features. Consequently, distance measures do not always correlate well with listeners evaluations of speech (natural or synthetic).

#### 1.4 Voice Quality

The quality of human voice is usually referred to as the total auditory impression the listener experiences upon hearing the speech of another talker. Most people have what is considered as a "normal" voice. Often, we meet people with creaky, breathy, rough or hoarse voices. In the literature on the human voice, vocal fry (creaky voice), modal and falsetto are often called "vocal registers," which are related to the production of voice pitch and the pitch range of an individual [Laver and Hanson, 1981]. A particular vocal register is characterized by a certain pattern of vocal fold vibration, with the vocal folds approximated in a similar way throughout a particular pitch range [Boone, 1971; Hollien, 1974]. Once this pitch range reaches its maximum limit, the vocal folds adjust to a new contour, producing an abrupt change in voice quality.

Severe breathiness, roughness and hoarseness are often considered voice disorders that may arise from laryngeal dysfunction. At the level of the larynx, these voice disorders are related to difficulties in the approximation of the vocal folds during phonations. It is generally hypothesized that laxity of vocal fold approximation generally produces an escape of air due to incomplete glottal closure, which is perceived as breathiness [Fant et al., 1985; Fant and Lin, 1988; Lee and Childers, 1989; Klatt and Klatt, 1990]. The excessive laryngeal tension during the vocal fold movement can lead to aperiodic vibrations of vocal folds, which is perceived as harshness or roughness [Coleman, 1960; Moore, 1975; Wendhal, 1963 and 1966]. The combination of escape of air and aperiodic vocal fold vibratory cycles is perceived as hoarseness [Yanagihara, 1967; Yumoto et al., 1984; Hiroka et al., 1984; Muta et al., 1988]. A literature review of various physiological, perceptual and acoustic characteristics for various voice types and disorders is given in Lee (1988). In general, a faulty muscular tension in various sites of the vocal mechanism, paralysis or development of cancer of vocal folds, growth of nodules on the vocal folds or abuse of the vocal mechanism result in a pathological condition of vocal mechanism. Such pathological conditions may permanently alter the physiology of the vocal mechanism. The symptoms of such pathologic conditions and the physiological alterations are often observed as permanent change or deviation from the speakers original "normal" voice.

A goal of speech scientists and phoneticians is to understand the vocal characteristics and the perceptual correlates of different types of voices, while speech pathologists and clinicians aspire to an understanding of various vocal disorders so that they may advance in their clinical practice and improve patients' care. A study of cause and effect relationships between the vocal excitation and the acoustic characteristics of the speech signal would certainly enhance their ability to identify, quantify and rank order the most likely laryngeal or the vocal characteristics that lead to the a specific

type of phonations and vocal disorders. We hypothesize that a flexible speech synthesizer may contribute to the advancement of speech science in such cases.

### 1.5 Research Goals and Plan

The currently available speech synthesizers, such as articulatory synthesizers [Maeda, 1982], Klatt's cascade/parallel formant synthesizer [Klatt, 1980; Klatt and Klatt, 1990], Holmes' all-parallel formant synthesizer [Holmes, 1983; Holmes et al., 1990], simple LPC synthesizers [Atal and Hanauer, 1971], multi-pulse LPC synthesizers [Atal and Remde, 1982], codebook excited LPC synthesizer [Schroeder and Atal, 1985], etc., can synthesize high-quality speech when the synthesizer parameters are properly controlled. However, the current implementations of these synthesizers have a limited parameter set and have a rigid synthesizer architecture. Hence, the current implementations of these synthesizers can be used only for a few types of experiments and applications. The first goal of this study was to develop a flexible speech synthesizer that can be used as a tool for conducting a wide variety of experiments that involve synthetic speech, including the experiments for improving the quality of synthetic speech. Such a flexible speech synthesizer would be an invaluable tool in several areas of speech research and in the development of synthetic speech applications. After comparing the pros and cons of the articulatory, LPC and formant synthesizers, we chose to develop a flexible formant synthesizer. The choice of the formant synthesizer was influenced by the requirements for the completion of the second goal, described in the next paragraph. The survey of historic and current formant synthesizers in the literature is given in Appendix A. The two currently popular formant synthesizers are Klatt's cascade/parallel formant synthesizer [Klatt, 1980; Klatt and Klatt, 1990], and Holmes' all-parallel formant synthesizer [Holmes, 1983; Holmes et al., 1990]. The advantages and disadvantages of both these

synthesizers are also discussed in Appendix A. For the reasons described in Appendix A, we decided to enhance Klatt's cascade/parallel formant synthesizer [Klatt, 1980] by incorporating flexibility in its parameter specification procedure, synthesis algorithm and the synthesizer architecture. The simple block diagram of the flexible formant synthesizer is shown in Figure 1-1.

The other goal of this study was to demonstrate that the formant synthesizer could be used to model various speech disorders caused by a laryngeal dysfunction. The traditional approach for understanding vocal disorders has been to analyze the speech data collected from human subjects and to obtain statistical relationships between the physiological and acoustical parameters, and the severity of various vocal disorders. A weakness of this approach is that the subject may inadvertently vary factors other than those he/she was instructed to manipulate. Furthermore, no two patients in a clinical study have exactly the same medical conditions or history, although they may exhibit similar disorders or pathologies. Consequently, a study of vocal disorders may be confounded by factors unknown to the researchers. Our approach is to employ a formant synthesizer to study the relationships between the acoustical parameters and various speech disorders. The advantage of using a formant synthesizer is that the source characteristics can be precisely varied independent of the vocal tract characteristics. Thus, the source characteristics can be precisely controlled and systematically varied to obtain synthetic speech with the desired vocal characteristics. Another advantage is that current implementations of the formant synthesizer are known to produce high quality, natural sounding speech [Childers and Wu, 1990; Klatt and Klatt, 1990; Holmes et al., 1990]. The listeners can perceptually evaluate the naturalness of the vocal characteristics under investigation through the listening tests. Using the formant synthesizer, it is possible to obtain the cause-and-effect relationships between the glottal flow characteristics and various vocal characteristics. Our approach is summarized in the block diagram in Figure 1-2.

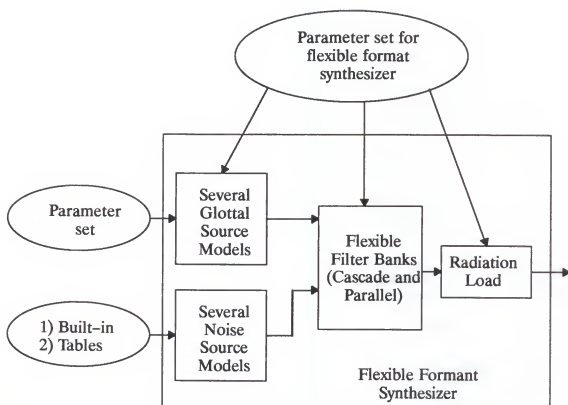


Figure 1-1: Simple block diagram for the flexible formant synthesizer



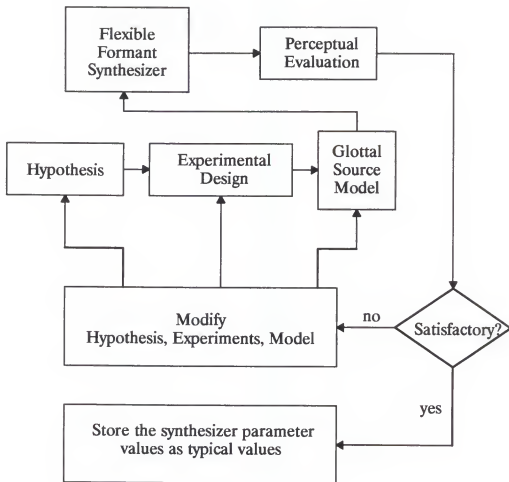


Figure 1-2: Block diagram for research approach to modeling speech disorders

For modeling various vocal characteristics through synthesis we developed the following:

- 1) a unified glottal source model that can specify acoustic parameters required for synthesizing various speech disorders, and
- 2) a procedure to systematically vary the parameters of this model in order to synthesize various speech disorders with variable severity.

The new glottal source model has been implemented and incorporated in the flexible formant synthesizer. The new glottal source model and the procedure for controlling its parameters will provide a basis for further study to find the quantitative descriptors (typical values of various acoustic parameters) of various speech disorders. These quantitative descriptors may eventually become universal among speech clinicians and researchers to characterize various speech disorders.

### 1.6 Description of the Following Chapters

In chapter 2 we first discuss the basic features of the flexible formant synthesizer and its parameter set. Although, many useful experiments can be performed using only the basic features of the flexible formant synthesizer, we modified the parameter set, parameter specification procedure, synthesis algorithms and the synthesizer architecture in order to expand the variety of experiments for which this synthesizer might be used. The new parameters, modified synthesis algorithms and synthesizer architecture are explained in chapter 3. We have incorporated several voicing source models in the flexible formant synthesizer. Our literature survey of research on speech disorders revealed that the existing glottal source models need modifications in order to synthesize high-quality normal speech and speech disorders. We developed a new unified glottal source model for synthesizing "high-quality" normal speech and various speech disorders. This glottal source model is described in chapter 4. In

chapter 5 we describe a procedure to systematically vary the parameters of this model in order to synthesize various voice disorders. In chapter 6 we describe the experiments conducted to assess the performance of this model in synthesizing/modeling various vocal disorders. In chapter 7 we summarize the results of this study and recommend future directions for further study.

## CHAPTER 2 A FLEXIBLE FORMANT SYNTHESIZER

### 2.1 Introduction

In this chapter, we present the theory upon which the formant synthesizers are based. The list of control and filter bank related parameters along with their minimum, typical and maximum values is presented along with the block diagram of the flexible formant synthesizer architecture. We discuss the basic features of the flexible formant synthesizer in detail. The chapter concludes with several synthesis strategies that may be employed for synthesizing various types of speech utterances using the flexible formant synthesizer.

### 2.2 Classification of Speech Sounds

On the basis of the excitation source, the speech sounds in American English are classified into three categories: voiced, unvoiced and mixed excitation sounds. A summary of the classification of various types of sounds in American English is shown in Table 2-I. The classification of the phonemes in American English is based upon both the excitation source and the configuration of the vocal tract during the production of each phoneme. A classification of American English phonemes is given in Figure 2-1.

Table 2-I  
Classification of speech sounds based upon the sound source

| Type of Sound Source | Place of Excitation         | Nature of Excitation                                              | Type of Sound                                             | Selected Examples                             |
|----------------------|-----------------------------|-------------------------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------|
| Voiced               | Vocal folds (Glottis)       | Quasi-periodic glottal pulses                                     | Vowels<br>Diphthongs<br>Semivowels<br>Nasals              | /i/, /a/<br>/ɔɪ/ /aʊ/<br>/w/, /j/<br>/m/, /n/ |
| Unvoiced             | Vocal tract                 | Aperiodic random noise                                            | Plosives<br>Fricatives<br>Affricates                      | /p/, /t/<br>/f/, /θ/<br>/tʃ/                  |
| Mixed                | Vocal tract and vocal folds | Aperiodic random noise modulated by quasi-periodic glottal pulses | Voiced Plosives<br>Voiced Fricatives<br>Voiced Affricates | /b/, /d/<br>/v/, /ð/<br>/dʒ/                  |

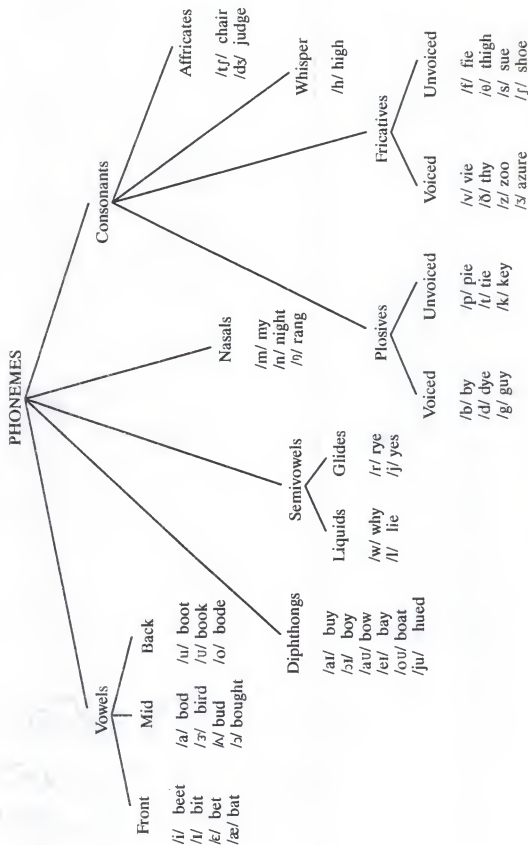


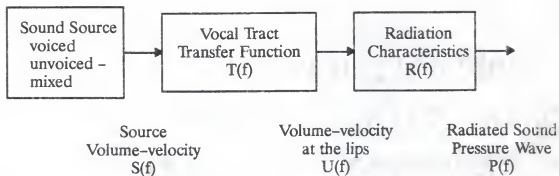
Figure 2-1: Phoneme classification in American English

## 2.3 Flexible Formant Synthesizer

The formant synthesizer design is based upon the acoustic theory of speech production presented by Fant (1960). The simple source-filter design of the formant synthesizer is summarized in Figure 2-2. According to this view, one or more sources of energy are activated by the buildup of lung pressure. Treating each sound source separately, as seen from Table 2-I, we may characterize it in the frequency domain by a source spectrum,  $S(f)$ , where “ $f$ ” is frequency in Hz. Each sound source excites the vocal tract which acts as a resonating system, analogous to an organ pipe. The acoustic theory of speech production assumes the vocal tract is a linear system and can be characterized in the frequency domain by a linear transfer function,  $T(f)$ , which is the ratio of volume velocity at the lips,  $U(f)$ , to the sound source input,  $S(f)$ . Production of each phoneme requires a different vocal tract configuration, and therefore, each phoneme has a separate transfer function,  $T(f)$ . The spectrum of the sound pressure,  $P(f)$ , that is typically recorded some distance from the lips of the talker, is related to the volume velocity at the lips,  $U(f)$ , by a radiation characteristic (load),  $R(f)$ , that describes the effects of directional sound propagation from the head. For speech synthesis the radiation load is typically kept constant for all phonemes.

### 2.3.1 Need for a Flexible Formant Synthesizer

Several formant synthesizer architectures have been implemented based upon the acoustic theory of speech production. A review of the past and currently popular formant synthesizers is given in Appendix A. The two most popular formant synthesizers are the cascade/parallel formant synthesizer designed by Klatt [Klatt, 1980] and the all-parallel formant synthesizer designed by Holmes [Holmes, 1983]. Both Klatt's and Holmes' synthesizers are capable of synthesizing highly intelligible speech when controlled by the proper synthesizer parameters. Therefore, these



$$P(f) = S(f) * T(f) * R(f)$$

Figure 2-2: A simple block diagram representing the acoustic theory of speech production



synthesizers are useful for speech synthesis in many commercial applications, as well as for some experiments in speech research. However, these synthesizers lack flexibility with respect to parameter specification and architecture. The limitations of these two synthesizers are discussed in Appendix A. This dissertation developed new software with flexibility in control parameter specification and architecture. This development was based on Klatt's cascade/parallel formant synthesizer for the following reasons:

- 1) It is popular among speech researchers.
- 2) Its configuration is more flexible than Holmes' all-parallel formant synthesizer. It has both the cascade and the parallel filter banks, and thus, allows a cascade/parallel synthesizer configuration and the all-parallel synthesizer configuration.

### 2.3.2 Development of a Flexible Formant Synthesizer

Our objective was to develop a flexible formant speech synthesizer that can synthesize intelligible and natural sounding speech. This synthesizer should also be flexible enough to be useful as a tool in conducting several types of experiments with synthetic speech. Such a flexible formant synthesizer should have:

- 1) a set of parameters whose values can be extracted from natural speech or specified from the parameter databases,
- 2) an architecture and synthesis algorithm that is capable of synthesizing speech with time and frequency domain characteristics as similar as possible to those of the natural speech sounds that the synthesizer is attempting to mimic, and
- 3) a capability to reconfigure the synthesis algorithm and synthesizer architecture through appropriate parameter specification.

It was simpler to write a new software package for the flexible formant synthesizer than to modify the software for Klatt's cascade/parallel formant synthesizer. We decided to implement the flexible formant synthesizer such that it simulates Klatt's

cascade/parallel formant synthesizer in its default configuration. After developing the basic formant synthesizer (Klatt's cascade/parallel formant synthesizer), we decided to enhance the basic formant synthesizer so that additional experiments to improve the quality of synthesized speech could be conducted. For these enhancements, we added several new parameters to the list of parameters of the basic formant synthesizer, and modified the synthesis algorithms and synthesizer architecture accordingly. For further expansion of the variety of experiments in which the flexible formant can be used, the flexible formant synthesizer may need further enhancements.

As a guideline for further enhancements of the flexible formant synthesizer a few comments are necessary. The decision about which acoustic parameters should be added to the flexible formant synthesizer's parameters list depends upon the types of experiments the researcher wishes to conduct. While modifying the flexible formant synthesizer implementation, these additional parameters should be included in the list of parameters if not previously present, and the synthesis algorithm and/or synthesizer architecture should be modified, if needed. Not all the experiments require addition of new parameters and/or modification of the synthesis algorithms and synthesizer architecture. Such experiments can be performed with the existing parameter set, synthesis algorithms and synthesizer architecture. In all these experiments, the experimenter should be able to specify the time and frequency domain characteristic he/she wants in the synthesized speech tokens by means of appropriate parameter specifications. The synthesizer architecture should be capable of producing the desired time and frequency domain characteristics in the synthetic speech from the input parameters. For example, if five formant frequencies and bandwidths are specified as input to a formant synthesizer then the spectrum of the synthesized speech token should display five peaks at specified frequencies and with specified widths.

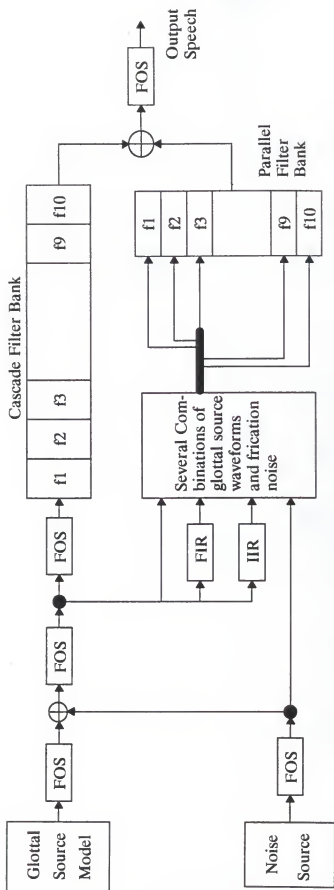
### 2.3.3 Synthesizer Block Diagram

The block diagram of the flexible formant synthesizer architecture is shown in Figure 2-3 and is discussed later.

### 2.3.4 List of Synthesizer's Parameters

Since the flexible formant synthesizer is an outgrowth of Klatt's cascade/parallel formant synthesizer [Klatt, 1980], the minimum, typical (default) and maximum values for the parameters common to Klatt's synthesizer have been retained for the flexible formant synthesizer. However, the symbols of the common parameters have in some cases been changed. In Table 2-II the control and filter banks (vocal tract) related parameters of the flexible formant synthesizer are listed along with their minimum, typical and maximum values of these parameters. Excluded from the table are the parameters of the glottal source models used with the flexible formant synthesizer. These parameters specify the shape of the glottal source pulses used as an excitation source in the synthesizer. The various glottal source models and their corresponding parameters are described in Appendix B.

Table 2-II lists a total of 62 control and filter bank related parameters. The flexible formant synthesizer parameters are more extensive than other synthesizers to improve the quality of synthesized speech and to allow the flexibility in using the synthesizer in a wide variety of speech experiments. The minimum, typical and maximum values of the parameters (source and vocal tract related parameters) given in these tables are typical for a male voice. These values will vary to approximate child and female voices. The "Formant Synthesizer User's Manual" explains in detail how the synthesizer parameters (control, vocal tract and source parameters) are specified as input to the flexible formant synthesizer.



FOS : First Order System (which may be a FIR filter, an IIR filter or a by-pass path)  
 f1 ... f10 : Second Order Systems (which may be a resonators, an anti-resonators or multipliers)

Figure 2-3: Configuration of the flexible formant synthesizer

Table 2-II

Table of control and filter bank parameters for the flexible formant synthesizer with minimum, typical and maximum values for a male voice

| # Parameter Name                                | Min. Value | Typical Value | Max. Value |
|-------------------------------------------------|------------|---------------|------------|
| 1) Sampling rate in Hz                          |            |               |            |
| sam_rat                                         | 5000.0     | 10000.0       | 20000.0    |
| 2) Frame size in terms of number of samples     |            |               |            |
| frame_size                                      | 0          | 50            | 256        |
| 3) Architecture type                            |            |               |            |
| arch_typ                                        | 1          | 3             | 3          |
| 4) Glottal source type                          |            |               |            |
| src_typ                                         | 1          | 7             | 9          |
| 5) Noise source type                            |            |               |            |
| nos_typ                                         | 1          | 1             | 2          |
| 6) Fundamental frequency in Hz                  |            |               |            |
| f0                                              | 0.0        | 0.0           | 500.0      |
| 7) Overall volume control in dB                 |            |               |            |
| g0                                              | 0.0        | 0.0           | 500.0      |
| 8) Voicing gain control in dB                   |            |               |            |
| av                                              | 0.0        | 0.0           | 80.0       |
| 9) Aspiration noise gain control in dB          |            |               |            |
| ah                                              | 0.0        | 0.0           | 80.0       |
| 10) Frication noise gain control in dB          |            |               |            |
| af                                              | 0.0        | 0.0           | 80.0       |
| 11) Amplitude of the first filter in dB         |            |               |            |
| a1                                              | 0.0        | 0.0           | 80.0       |
| 12) Bandwidth of the first filter in Hz         |            |               |            |
| b1                                              | 40.0       | 50.0          | 500.0      |
| 13) Center frequency of the first filter in Hz  |            |               |            |
| f1                                              | 150.0      | 450.0         | 900.0      |
| 14) Amplitude of the second filter in dB        |            |               |            |
| a2                                              | 0.0        | 0.0           | 80.0       |
| 15) Bandwidth of the second filter in Hz        |            |               |            |
| b2                                              | 40.0       | 70.0          | 500.0      |
| 16) Center frequency of the second filter in Hz |            |               |            |
| f2                                              | 500.0      | 1450.0        | 2500.0     |
| 17) Amplitude of the third filter in dB         |            |               |            |
| a3                                              | 0.0        | 0.0           | 80.0       |

Table 2-II ... Continued

| #   | Parameter Name                               | Min. Value | Typical Value | Max. Value |
|-----|----------------------------------------------|------------|---------------|------------|
| 18) | Bandwidth of the third filter in Hz          |            |               |            |
| b3  |                                              | 40.0       | 110.0         | 500.0      |
| 19) | Center frequency of the third filter in Hz   |            |               |            |
| f3  |                                              | 1300.0     | 2450.0        | 3500.0     |
| 20) | Amplitude of the fourth filter in dB         |            |               |            |
| a4  |                                              | 0.0        | 0.0           | 80.0       |
| 21) | Bandwidth of the fourth filter in Hz         |            |               |            |
| b4  |                                              | 100.0      | 250.0         | 500.0      |
| 22) | Center frequency of the fourth filter in Hz  |            |               |            |
| f4  |                                              | 2500.0     | 3300.0        | 4500.0     |
| 23) | Amplitude of the fifth filter in dB          |            |               |            |
| a5  |                                              | 0.0        | 0.0           | 80.0       |
| 24) | Bandwidth of the fifth filter in Hz          |            |               |            |
| b5  |                                              | 150.0      | 200.0         | 700.0      |
| 25) | Center frequency of the fifth filter in Hz   |            |               |            |
| f5  |                                              | 3500.0     | 3750.0        | 4900.0     |
| 26) | Amplitude of the sixth filter in dB          |            |               |            |
| a6  |                                              | 0.0        | 0.0           | 80.0       |
| 27) | Bandwidth of the sixth filter in Hz          |            |               |            |
| b6  |                                              | 200.0      | 1000.0        | 2000.0     |
| 28) | Center frequency of the sixth filter in Hz   |            |               |            |
| f6  |                                              | 4000.0     | 4900.0        | 4999.0     |
| 29) | Amplitude of the seventh filter in dB        |            |               |            |
| a7  |                                              | 0.0        | 0.0           | 80.0       |
| 30) | Bandwidth of the seventh filter in Hz        |            |               |            |
| b7  |                                              | 50.0       | 100.0         | 500.0      |
| 31) | Center frequency of the seventh filter in Hz |            |               |            |
| f7  |                                              | 200.0      | 250.0         | 700.0      |
| 32) | Amplitude of the eighth filter in dB         |            |               |            |
| a8  |                                              | 0.0        | 0.0           | 80.0       |
| 33) | Bandwidth of the eighth filter in Hz         |            |               |            |
| b8  |                                              | 50.0       | 100.0         | 500.0      |
| 34) | Center frequency of the eighth filter in Hz  |            |               |            |
| f8  |                                              | 200.0      | 250.0         | 500.0      |
| 35) | Amplitude of the ninth filter in dB          |            |               |            |
| a9  |                                              | 0.0        | 0.0           | 80.0       |

Table 2-II ... Continued

| # Parameter Name                                                                           | Min. Value | Typical Value | Max. Value |
|--------------------------------------------------------------------------------------------|------------|---------------|------------|
| 36) Bandwidth of the ninth filter in Hz                                                    |            |               |            |
| b9                                                                                         | 0.0        | 0.0           | 0.0        |
| 37) Center frequency of the ninth filter in Hz                                             |            |               |            |
| f9                                                                                         | 0.0        | 0.0           | 0.0        |
| 38) Amplitude of the tenth filter in dB                                                    |            |               |            |
| a10                                                                                        | 0.0        | 0.0           | 80.0       |
| 39) Bandwidth of the tenth filter in Hz                                                    |            |               |            |
| b10                                                                                        | 0.0        | 0.0           | 0.0        |
| 40) Center frequency of the tenth filter in Hz                                             |            |               |            |
| f10                                                                                        | 0.0        | 0.0           | 0.0        |
| 41) Type and coefficient of the first order filter in the cascade filter bank              |            |               |            |
| c_filt                                                                                     | -1.0       | 0.0           | 1.0        |
| 42) Type and coefficient of the first order filter with the glottal source model           |            |               |            |
| g_filt                                                                                     | -1.0       | -1.0          | 1.0        |
| 43) Type and coefficient of the first order filter with the noise source model             |            |               |            |
| n_filt                                                                                     | -1.0       | 0.0           | 1.0        |
| 44) Type and coefficient of the first order filter at the output                           |            |               |            |
| o_filt                                                                                     | -1.0       | 0.0           | 1.0        |
| 45) Coefficient of the first order highpass filter in the parallel filter bank             |            |               |            |
| ph_filt                                                                                    | -1.0       | -1.0          | 1.0        |
| 46) Coefficient of the first order lowpass filter in the parallel filter bank              |            |               |            |
| pl_filt                                                                                    | -1.0       | 0.99          | 1.0        |
| 47) Type and coefficient of the first order filter with the source models                  |            |               |            |
| u_filt                                                                                     | -1.0       | 0.0           | 1.0        |
| 48) Exponent for the plosive exponential                                                   |            |               |            |
| step_size                                                                                  | 0.0        | 0.5           | 1.0        |
| 49) Flag set for automatic determination of scale factors for parallel filter bank filters |            |               |            |
| PLUS_MINUS                                                                                 | 0          | 0             | 1          |
| 50) Starting frame number to begin synthesis                                               |            |               |            |
| start_frame                                                                                | 0          | 0             | 1000       |
| 51) Total number of frames to be synthesized                                               |            |               |            |
| tot_frames                                                                                 | 0          | 0             | 1000       |
| 52) Starting time to begin synthesis                                                       |            |               |            |
| start_dur                                                                                  | 0.0        | 0.0           | 2.0        |
| 53) Total duration of utterance to be synthesized                                          |            |               |            |
| tot_dur                                                                                    | 0.0        | 0.0           | 2.0        |

Table 2-II ... Continued

| # Parameter Name                                                                   | Min. Value | Typical Value | Max. Value |
|------------------------------------------------------------------------------------|------------|---------------|------------|
| 54) Use start_frame and tot_frames instead of start_dur and tot_dur                |            |               |            |
| FRAMES                                                                             | 0          | 1             | 1          |
| 55) Synthesis mode is pitch-synchronous                                            |            |               |            |
| PITCH_SYNC                                                                         | 0          | 0             | 1          |
| 56) Simulation of source-tract interaction by abrupt change in "f1" and "b1"       |            |               |            |
| ST_FRAME                                                                           | 0          | 0             | 1          |
| 57) Simulation of source-tract interaction by smooth change in "f1" and "b1"       |            |               |            |
| ST_SMP                                                                             | 0          | 0             | 1          |
| 58) fractional change in the value of "f1" during the open-phase                   |            |               |            |
| for_freq                                                                           | 0.0        | 1.2           | 2.0        |
| 59) fractional change in the value of "b1" during the open-phase                   |            |               |            |
| for_bw                                                                             | 0.0        | 1.2           | 2.0        |
| 60) open-phase duration as a fraction of pitch-period                              |            |               |            |
| op                                                                                 | 0.0        | 0.8           | 1.0        |
| 61) Threshold for ratio of initial condition responses of the cascade filter bank  |            |               |            |
| tran_cas                                                                           | 0.0        | 100.0         | 1000.0     |
| 62) Threshold for ratio of initial condition responses of the parallel filter bank |            |               |            |
| tran_par                                                                           | 0.0        | 100.0         | 1000.0     |



### 2.3.5 Synthesizer Software/Hardware and Flowchart

Some highlights of the synthesizer software and a brief description of the hardware requirements are given in Appendix C, along with a flowchart of the synthesizer algorithm.

### 2.3.6 Assessing the Performance of the Flexible Formant Synthesizer

In our literature survey of various speech synthesis systems, we observed that formal listening tests were conducted when the performance of various rule-based systems or the performance of various analysis-synthesis systems were being evaluated. We did not find a study describing the quantitative assessment of the performance of a synthesizer. The reason for not using quantitative methods may be due to the lack of their correlation with the listeners perceptual evaluations. Quantitative assessment methods were used in a few applications of LPC synthesizers, such as variable rate parameter transmission vocoders, etc. The formal listening tests and the quantitative methods have not been used to evaluate the performance of the recently developed synthesizers by Klatt and Klatt (1990) and Holmes et al. (1990).

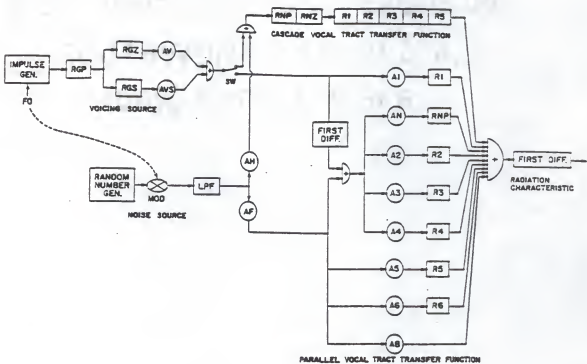
Klatt (1980 and 1987), Klatt and Klatt (1990), Holmes (1973 and 1983) and Holmes et al. (1990) have conducted informal listening tests and visual comparisons between the synthesized speech spectrograms and natural speech spectrograms to evaluate the performance of their speech synthesizers. The informal listening tests were used to check if the synthesized speech was intelligible and natural sounding when "good" acoustic parameters were given as the input. The speech spectra and spectrograms were used to compare the similarity between synthetic speech signals and natural speech signals at acoustically critical regions, such as vowel-vowel transitions, consonant-vowel transitions, etc.

## 2.4 Basic Features of the Flexible Formant Synthesizer

The typical (default) values of the flexible formant synthesizer parameters give Klatt's cascade/parallel formant synthesizer [Klatt, 1980]. Figure 2-4a shows the block diagram of Klatt's cascade/parallel formant synthesizer as described in Klatt (1980). Since we refer to the Holmes (1983) all-parallel formant synthesizer from time-to-time, Figure 2-4b illustrates this synthesizer's block diagram. The following sub-sections describe the basic features and parameters of the flexible formant synthesizer.

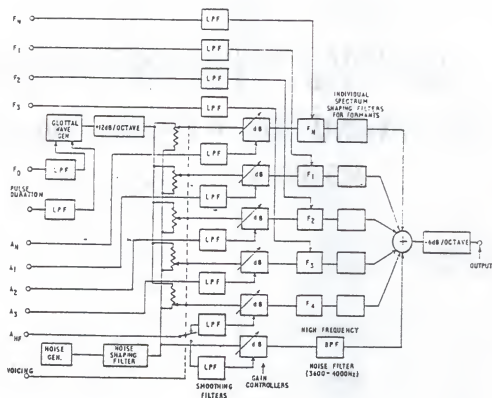
### 2.4.1 Digital Implementation in the Time Domain

The design of the formant synthesizer is often described in the frequency domain. However, the design can be recast in the time domain, where many synthesizer characteristics are more easily understood and implemented. For example, the time domain design requires fewer computations using difference equations as compared to calculating FFTs (Fast Fourier Transforms) and IFFTs (Inverse Fourier Transforms). One might suppose that the frequency response of the glottal source model, filter bank and radiation load might be calculated and stored. Then interpolation and decimation techniques might be used to adjust the frequency response to new values. However, the shape of the frequency response is not changed by this approach. Another design approach might be to implement the filter banks with the overlap-add or overlap-save methods [Oppenheim and Schaffer, 1975]. A problem with this approach is truncating the impulse response of the filter banks, especially when the frame size is small. Also, the speech waveform obtained by concatenating the time domain waveform generated for adjacent frames may not be smooth or continuous at the frame boundaries. Such discontinuities can cause distortions in the speech waveform that are audible.



[a]

Figure 2-4: Block diagram of  
 a) Klatt's cascade/parallel formant synthesizer  
 b) Holmes' all-parallel formant synthesizer



[b]

Figure 2-4: Continued

### 2.4.2 Control Parameters for the Flexible Formant Synthesizer

The control parameters of the formant synthesizer are the sampling rate of the speech waveform, (smp\_rat), the frame size. (frame\_size), the starting frame, (start\_frame), the total number of frames to be synthesized. (tot\_frames), the starting instant of synthesis (in seconds), (start\_dur), and the total duration of the synthesis (in seconds), (tot\_dur).

#### 2.4.2.1 Waveform sampling rate

Most of the speech sound energy is contained in the frequency band between 80 Hz and 8000 Hz [Dunn and White, 1940]. However, intelligibility tests of bandpass filtered speech indicate that intelligibility is not measurably changed if the energy in the frequency band above 5000 Hz is removed [French and Steinberg, 1947]. Speech that is low-pass filtered in this manner sounds quite natural. The typical sampling rate (samples per second) ranges from 8000 Hz to 20000 Hz. The sampling rate may be specified by the parameter "sam-rat."

#### 2.4.2.2 Parameter update rate

The synthesizer parameter update rate (in seconds) during the synthesis of an utterance is determined by both the sampling rate and the speech frame size. A speech frame is the portion of the speech signal during which the time and frequency domain characteristics are considered to remain stationary. The parameter "frame\_size" specifies the number of samples to be generated in a frame. Normally, the control parameters are updated at every 5 msec interval. This rate is frequent enough to mimic most sounds including those with rapid formant transitions and brief plosive bursts [Klatt, 1980]. The value of the parameter "frame\_size" is normally kept constant during the synthesis. In some cases, however, it is more appropriate to vary

the frame size during the synthesis. Such cases arise when the values of the variable parameters correspond to variable durations of the portions of an utterance being synthesized. For example, the values of the synthesis parameters such as formant frequencies and bandwidths can be obtained by using the pitch-synchronous analysis techniques during the voiced portions of an utterance and by using the fixed frame analysis techniques during the unvoiced/silence portions of an utterance. During resynthesis of such an utterance, if the parameter "PITCH\_SYNC" (flag) is set (value set equal to one), the voiced portions are synthesized by the pitch-synchronous synthesis method and unvoiced portions are synthesized by the fixed-frame synthesis method. The frame size for the fixed-frame synthesis method is kept constant as specified by the "frame\_size" parameter. The frame size for the pitch-synchronous synthesis method is modified automatically (by the flexible formant synthesizer) on a period-by-period basis to be equal to the number of samples in each pitch-period.

#### 2.4.2.3 Duration of an utterance

The total duration (in seconds) of an utterance being synthesized is a product of the sampling interval and the total number of samples to be synthesized. The total number of samples to be synthesized depends upon the total number of frames to be synthesized and size of each frame. The total number of frames to be synthesized is specified by the parameter "tot\_frames." If the parameter "frame\_size" is kept constant then the total number of samples to be synthesized is a product of the "frame\_size" and "tot\_frames." If the parameter "frame\_size" is variable then the total number of samples to be synthesized is the sum of the number of samples in all the frames to be synthesized. The value of the "start\_frame" parameter is kept zero when the utterance has to be synthesized from its beginning. When only a portion of an utterance is to be synthesized, the parameter "start\_frame" specifies the starting frame to begin synthesis. In that case, the parameter "tot\_frames" is interpreted as the total

number of frames to be synthesized after the starting frame. The utterance to be synthesized should have at least `start_frame + tot_frames` number of samples in each of the variable parameter tracks. A variable parameter track is an array for values for that parameter in successive frames of an utterance.

Instead of specifying the duration of an utterance in terms of the number of frames, it can be specified directly by its total duration in seconds by the parameter `"tot_dur."` Then the total number of samples to be synthesized is a product of the `"sam_rat"` and `"tot_dur"` parameters. If the parameter `"frame_size"` is a constant, then the total number of frames to be synthesized is obtained by dividing the total number of samples by the value of the parameter `"frame_size."` If the parameter `"frame_size"` is a variable (i.e., specified by a variable parameter track), then the total number of frames to be synthesized is equal to the maximum value of the integer number of consecutive frames, whose total duration adds up to the duration specified by the parameter `"tot_dur."` The parameter `"start_dur"` can be used just like `"start_frame"` to synthesize only a portion of the utterance, except that the value of the beginning instant of synthesis of an utterance should be specified in seconds. The parameter `"FRAMES"` is used as a flag to indicate which method for calculation of the total duration of an utterance (number of samples) should be used. When this flag is set, the parameters `"start_frame"` and `"tot_frames"` determine the total duration of the utterance to be synthesized. Otherwise the parameters `"start_dur"` and `"tot_dur"` determine the total duration.

When synthesizing a sustained utterance, such as a sustained /a/ vowel, the flag `"FRAMES"` should be set, the parameter `"tot_frames"` should be used to specify the total number of frames to be synthesized and the parameter `"start_frame"` should be equal to zero. Or the parameter `"tot_dur"` should be used to specify the total duration of the vowel to be synthesized and the parameter `"start_dur"` should be equal to zero. When synthesizing an utterance for which parameter tracks are specified, the total

number of frames that can be synthesized is automatically set equal to the length (number of samples) of the variable parameter track(s) by the flexible formant synthesizer. The length of every parameter track to be used for the synthesis of an utterance should be the same. However, if the lengths are unequal, a warning message is displayed and the total number of frames that can be synthesized is set equal to the length of the shortest variable parameter track. The value of the parameter “tot\_frames” should be less than or equal to the total number of frames that can be synthesized. The value of the parameter “tot\_dur” should be less than the original duration of the utterance from which the parameter tracks were obtained. The parameter “start\_frame” or “start\_dur” should be equal to zero when synthesizing a complete utterance. When only a portion of the utterance has to be synthesized, the parameter “start\_frame” or “start\_dur” can be nonzero.

#### 2.4.3 Glottal Source Models

A glottal source model is used as a voicing source for synthesizing voiced sounds, such as vowels (/i/, /ɪ/, /ɛ/, /æ/, /a/, /ɜ/, /ʌ/, /ɔ/, /u/, /ʊ/ and /o/), semi-vowels (/w/, /l/, /r/ and /j/), diphthongs (/aɪ/, /ɔɪ/, /aʊ/, /eɪ/ and /ju/) and nasals (/m/, /n/ and /ŋ/). For these types of sounds, the sound source is at the glottis. A glottal source model simulates the production of glottal flow pulses. In the flexible formant synthesizer we provide several glottal source models. The parameter “src\_tpy” specifies the glottal source model to be used as a voicing source during the synthesis. We have provided both the parametric and non-parametric glottal source models. The parametric source models, based on a set of the input parameters, generate a stylized glottal source waveform that closely resembles the shape of the glottal flow pulses produced at the glottis. A review of some of the recent parametric glottal source models is given in Fujisaki and Ljungqvist (1986), Ananthapadmanabha (1984) and Childers and Wu (1990) have shown that highly sophisticated glottal source



models can improve the quality of synthesized speech. The specifications for the glottal source models that are important for high quality speech are given in Childers and Wu (1990). The non-parametric glottal source models are the amplitude-time waveforms of pulse(s) which closely resemble the glottal flow pulses. Generally, these waveforms are obtained by inverse filtering of speech by a typical talker [Holmes, 1973]. The non-parametric glottal source waveforms can also be obtained by using Sondhi's reflectionless tube or Rothenberg's mask while the speech utterances are spoken by a typical talker. Childers and Wu (1990) and Holmes (1973) have shown that the waveforms obtained by these methods retain the spectral fine details and the time structure present in the glottal flow waveforms, and therefore, the synthetic speech using these waveforms sound natural.

In the flexible formant synthesizer, the parameter "f0" specifies the fundamental frequency, which is the rate of repetition of glottal source pulses during voiced sounds. The parameter "av" specifies the value (in dB) of either the energy, power or peak amplitude in each glottal source pulse, which is dependent upon the value of another parameter, "typ\_gain." (The parameter "typ\_gain" is considered as a glottal source model parameter, and hence, not included in Table 2-II.) A large value of the parameter "av" (60 dB) indicates a strong voicing source, while a zero value "turns-off" the voicing source. A glottal source pulse is generated when the values of both the "av" and "f0" parameters are nonzero. The glottal source pulses are generated at the rate specified by the fundamental frequency parameter "f0" for the consecutive frames with nonzero values of both the "av" and "f0" parameters. These parameters do not change the shape of the glottal source pulses. The shape of the glottal source pulses is determined by the parameters of the glottal source model selected for synthesis. Various glottal source models are described in Appendix B.

#### 2.4.4 Noise Source

A white-noise source is normally used for both aspiration and frication in the synthesis of aspiration and whisper (/h/), plosives (/p/, /t/, /k/, /b/, /d/ and /g/) and fricatives (/f/, /θ/, /s/, /ʃ/, /v/, /θ/, /z/ and /ʒ/). For aspiration and whisper sounds the source is located at the glottis. For plosives and fricatives the noise source is located at an occlusion or a narrow constriction in the vocal tract, respectively. In the flexible formant synthesizer we provide two types of noise sources. The parameter "nos\_typ" specifies the noise source to be used during the synthesis. The noise source can be a random number sequence generated by a built-in random number generator or specified through an external random number table. A random number sequence with Gaussian distribution and white-noise frequency characteristics is commonly used as a noise source [Klatt, 1980; Holmes, 1983]. The problem with using random number sequences as a noise source is that short random number sequences in an utterance may not approximate a Gaussian distribution and/or white-noise characteristics and the listeners may misinterpret the sound generated by the "colored-noise" (narrowband, short-duration noise bursts). To alleviate this problem, in the flexible formant synthesizer, the noise source is left "turned-on" for the entire duration of the utterance being synthesized. However, the power of the random number sequence during the synthesis of voiced sounds should be attenuated to less than 20 dB.

The random numbers generated by the built-in random number generator, are scaled by an appropriate factor so that the power for a long random number sequence is unity. It is also preferable to have a random number sequence specified through the external random number tables to be a unit power sequence. The parameter "ah" specifies the gain of the aspiration noise source (in dB). The parameter "af" specifies the gain of the frication noise source (in dB). A high value of the parameter "ah" (60 dB) indicates a strong aspiration source. A high value of parameter "af" (60 dB)

indicates a strong frication source. We have also assumed that the random number sequence is wide sense stationary and the power in any portion of the sequence remains constant. Accordingly, the random number sequence is scaled by the value of the parameter “ah” or “af” on a frame-by-frame basis to adjust the power in the random number sequence for each frame as specified by their parameter tracks. A single noise source, i.e., a single random number sequence, is used as the noise source for both the aspiration and frication sounds. The random number sequence is scaled by the value of the “ah” or the “af” parameter, whichever is larger. The “ah” and “af” parameters do not affect the spectrum of the noise source. The noise sources that can be specified in the flexible formant synthesizer are described in Appendix B.

#### 2.4.5 Modulation of the Noise Source

A good model for an excitation source for synthesizing mixed excitation sounds (voiced plosives /b/, /d/ and /g/, and voiced fricatives, /v/, /ð/, /z/, /ʒ/) is a voicing source used in conjunction with a noise source that is modulated by an amplitude-time waveform [Klatt, 1980]. The aspirated and whisper sounds can be synthesized using the amplitude modulated noise source alone. For synthesizing these sounds (mixed, aspirated and whisper), the random number sequence is amplitude-modulated, pitch-synchronously, by an amplitude-time waveform whose duration is equal to the pitch-period. Amplitude-modulation simulates the effect of the vibrating vocal folds on the steady air flow from the lungs during the production of these sounds. In the flexible formant synthesizer, the amplitude-modulation waveform has three parts for each glottal source pulse. The parameter “amp1” specifies the amplitude of the first and the third parts and the parameter “amp2” specifies the amplitude of the second part. The parameters “offset” and “dur” specify the duration of the first and the second parts, respectively. The duration of the third part is given by (pitch-period – (offset + dur)). Due to the amplitude-modulation, the intensity of the random number

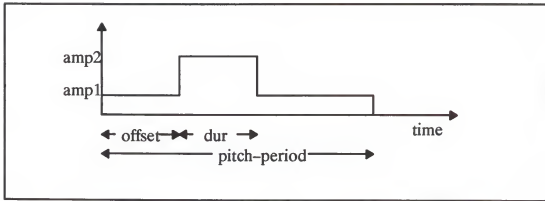
sequence will be time-varying, and hence, the flow of the noise power perceptible to the listener will also be time-varying. These specifications can simulate the amplitude-time waveform used for amplitude modulation of the noise source by Klatt (1980) and by Lee and Childers (1989). Since the amplitude-modulation waveform simulates the effect of the vocal folds (voicing source) on the noise source, the parameters "amp1," "amp2," "offset" and "dur" are considered as glottal source model parameters in the flexible formant synthesizer. Figure 2-5 describes the amplitude-modulation waveform and shows its effect on the noise source.

#### 2.4.6 Frequency-Shaping of the Noise Source

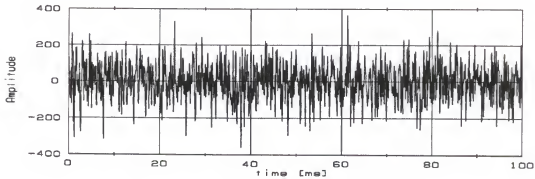
The spectrum of the noise source used for synthesis should be approximately flat [Stevens, 1971]. Holmes (1983) and other synthesizer implementations have used a noise source with a flat spectrum. Klatt (1980) has argued that the noise source with a flat spectrum simulates a constant pressure source and does not simulate the volume-velocity at the constriction. Considering the acoustic impedance of the front cavity and assuming that it is largely inductive, the volume-velocity at the constriction can be obtained by integration of the noise source. In Klatt's cascade/parallel formant synthesizer, a first order IIR (Infinite Impulse Response) filter is used as an approximation to the integration of the noise source (random number sequence). In the flexible formant synthesizer, filtering of the random number sequence from the noise source is optional. The spectra of the white-noise, lowpass filtered noise source and highpass filtered noise source are shown in Figure 2-6.

#### 2.4.7 First Order Systems

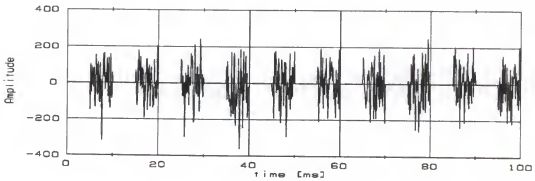
In Klatt's cascade/parallel formant synthesizer, the first order FIR (Finite Impulse Response) filters and the first order IIR (Infinite Impulse Response) filters are used to modify the spectra of the glottal source pulses, noise source and radiation



[a]



[b]



[c]

Figure 2-5: Noise source

a) Amplitude-modulation waveform

b) Noise source

c) Amplitude-modulated noise source

(amp1 = 0.0, amp2 = 1.0, offset = 0.5 and dur = 0.5)

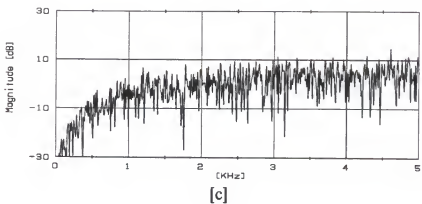
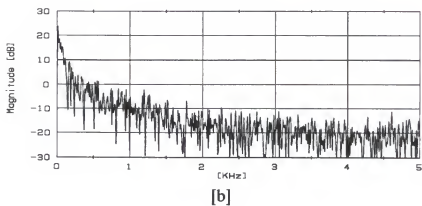
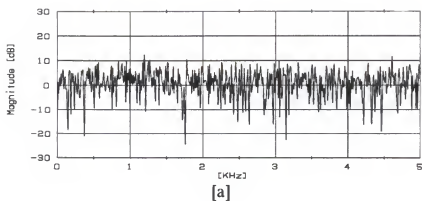


Figure 2-6: Frequency domain characteristics of noise source  
a) Spectrum of white-noise  
b) Spectrum of lowpass filtered noise source  
c) Spectrum of highpass filtered noise source

load. With the variety of glottal source models and the noise sources implemented in the flexible formant synthesizer, the user should have flexibility to select the type of filter used to modify the spectra of the glottal source pulses and the noise source. For example, if a differentiated glottal flow waveform obtained by inverse filtering of natural speech is used as a glottal source waveform then a first order FIR filter in series with the glottal source model may not be used during the resynthesis of that utterance. As another example, the user may want to study the effect of lowpass filtered noise, highpass filtered noise or unfiltered noise on the "quality" of synthesized breathy sounds. In the flexible formant synthesizer, we have created FOSs (First Order Systems) that can be used as a first order IIR filter, as a first order FIR filter or as a by-pass path with unity gain.

Let the input to a FOS be represented by  $x(n)$  and the output from the FOS be represented by  $y(n)$ . Let the filter coefficient 'a' have a "real" value. The difference equation for a FIR filter is given by

$$y(n) = x(n) + a \cdot x(n-1)$$

A "zero" is created at  $z = |a|$  if  $-1 \leq a < 0$ . A FIR filter with a "real zero" on the positive side of the X axis in the z-plane gives a magnitude frequency response of a highpass filter. The difference equation for an IIR filter is given by

$$y(n) = x(n) + a \cdot y(n-1)$$

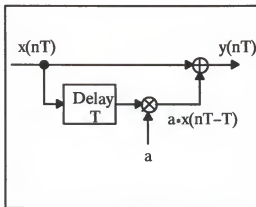
A "pole" is created at  $z = |a|$  if  $0 < a < 1$ . A first order IIR filter with a "real zero" on the positive side of the X axis in the z-plane gives a magnitude frequency response of a lowpass filter. Thus, by specifying a positive value to the filter coefficient 'a' the FOS can be configured to simulate a IIR filter, and by specifying a negative value of the filter coefficient the FOS can be configured to simulate an FIR filter. If  $a = 0$ , the FOS simulates a by-pass path with unity gain. This method is versatile for specifying both the type of the first order filter (a first order FIR filter, a first order IIR filter or a by-pass path) and the filter coefficient through a single parameter. The

first order IIR filter is normalized to 0 dB at dc. For the first order IIR filter, the filter coefficient 'a' should be specified to be less than 1, since specifying a  $\geq 1$  causes the filter to become unstable. The block diagram, the impulse response and magnitude frequency response of a first order FIR filter and a IIR filter are shown in Figure 2-7.

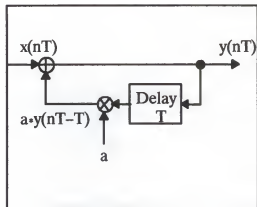
#### 2.4.8 Filter Banks

Propagation of the volume-velocity waveform through the vocal tract and the nasal tract depends upon the transfer function,  $T(f)$ . The transfer function,  $T(f)$ , relates the spectrum of the source volume-velocity to the spectrum of the volume-velocity at the lips. A formant synthesizer employs the cascade and/or the parallel filter bank(s) to simulate the magnitude frequency response of the transfer function,  $T(f)$ . The advantage of the formant synthesizer is that the user can directly specify the values of the formant and the anti-formant frequency and bandwidth as parameters of the synthesizer and observe the peaks (formants) and notches (anti-formants) with appropriate widths (bandwidths) in the magnitude frequency response of the filter bank. The formants are high-energy frequency-bands in speech sounds. They represent the resonance (build-up of energy) in the vocal and nasal tracts. The anti-formants are the low-energy frequency-bands in speech sounds. They represent the anti-resonances (loss of energy) in the vocal and nasal tracts. The formants and anti-formants can be observed in the spectra of speech sounds and the variations in their values (formant and anti-formant tracks) in an utterance can be observed in the speech spectrogram. Figure 2-8a and b show the speech signal and the spectrogram of a sentence "We were away a year ago." The formant tracks are observed as the dark bands in the spectrogram. The formant frequency tracks for this sentence are shown in Figure 2-8c. The formant and anti-formant frequencies and bandwidths are known to be perceptually significant for discriminating phonemes. Therefore, different phonemes can be synthesized by specifying different values for the formant

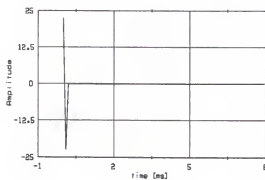




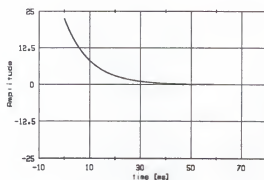
[a]



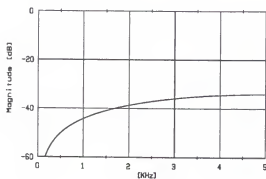
[b]



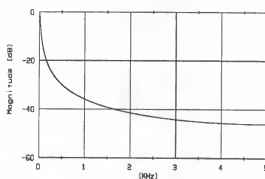
[c]



[d]



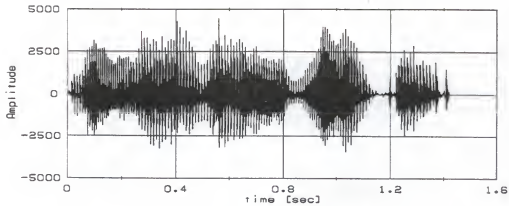
[e]



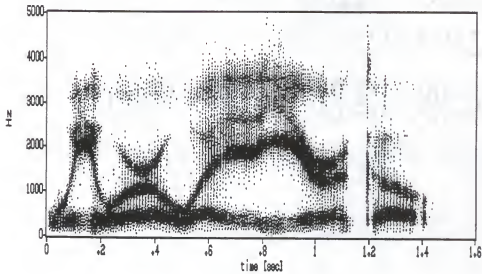
[f]

Figure 2-7: Characteristics for first order FIR and IIR filters

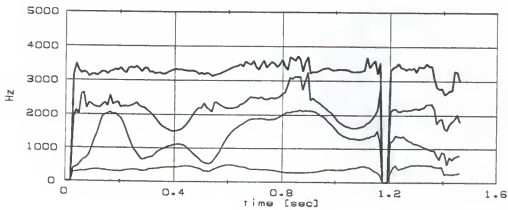
- a) Block diagram of a first order FIR filter
- b) Block diagram of a first order IIR filter
- c) Impulse response of a first order FIR filter
- d) Impulse response of a first order IIR filter
- e) Magnitude frequency response of a first order FIR filter
- f) Magnitude frequency response of a first order IIR filter



[a]



[b]



[c]

Figure 2-8: Sentence "We were away a year ago."  
 a) Speech signal  
 b) Spectrogram of speech signal  
 c) Formant frequency tracks

and anti-formant frequency and bandwidth parameters. The cascade and the parallel filter banks in the formant synthesizer use the second order resonators to simulate formants and the second order anti-resonators to simulate anti-formants.

#### 2.4.8.1 Digital resonator

Two parameters are used to specify the input-output characteristics of a digital resonator, the resonance (center or formant) frequency 'F' and the resonance (formant) bandwidth 'BW' (both in Hz). Samples of the output sequence of a digital resonator,  $y(nT)$ , are computed from the input sequence,  $x(n)$ , using the equation

$$y(nT) = A \cdot x(nT) + B \cdot y(nT-T) + C \cdot y(nT-2T)$$

where  $y(nT-T)$  and  $y(nT-2T)$  are the previous two samples of the output sequence  $y(nT)$ . 'T' is the sampling (time) interval between two samples and is given by the multiplicative inverse of "sam\_rat." The block diagram, impulse response and magnitude response of a second order digital resonator are given in Figure 2-9. The resonator's coefficients 'A', 'B' and 'C' are related to the resonance frequency 'F' and bandwidth 'BW' of a resonator by the impulse invariant transform [Gold and Rabiner, 1968].

$$C = e^{-2\pi BW T}$$

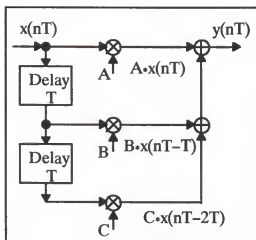
$$B = 2e^{-\pi BW T} \cos(2\pi FT)$$

$$A = 1 - B - C$$

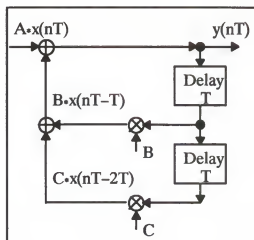
The transfer function of a second order digital resonator is given by

$$T(z) = \frac{A}{1 - Bz^{-1} - Cz^{-2}}$$

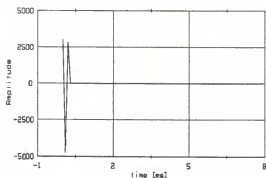
where  $z = e^{j2\pi fT}$  for obtaining the frequency response, 'j' is an imaginary number corresponding to the square root of -1, and 'f' is frequency in Hz, with a range from 0 to half the sampling rate. The resonator's coefficient 'A' insures that the magnitude response at dc is zero dB, i.e., the dc air flow passes unimpeded.



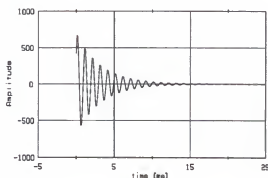
[a]



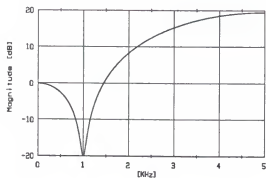
[b]



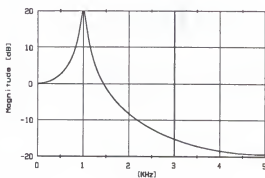
[c]



[d]



[e]



[f]

Figure 2-9: Characteristics for second order FIR and IIR filters  
 a) Block diagram of a second order FIR filter  
 b) Block diagram of a second order IIR filter  
 c) Impulse response of a second order FIR filter  
 d) Impulse response of a second order IIR filter  
 e) Magnitude frequency response of a second order FIR filter  
 f) Magnitude frequency response of a second order IIR filter

### 2.4.8.2 Digital anti-resonator

Two parameters are used to specify the input-output characteristics of a digital anti-resonator, the anti-resonance (center or anti-formant) frequency 'F' and anti-resonance (anti-formant) bandwidth 'BW' (both in Hz). Samples of the output sequence of a digital anti-resonator,  $y(nT)$ , are computed from the input sequence,  $x(n)$ , using the equation

$$y(nT) = A' \cdot x(nT) + B' \cdot x(nT-T) + C' \cdot x(nT-2T)$$

where  $x(nT-T)$  and  $x(nT-2T)$  are the previous two samples of the input sequence  $x(nT)$ . The anti-resonator's coefficients  $A'$ ,  $B'$  and  $C'$  are obtained as follows:

$$C' = -C/A$$

$$B' = -B/A$$

and

$$A' = 1.0/A$$

where 'A', 'B' and 'C' are obtained as shown earlier. The transfer function of a second order digital anti-resonator is given by

$$T(z) = A' + B'z^{-1} + C'z^{-2}$$

The block diagram, impulse response and magnitude response of digital second order resonator and anti-resonator are given in Figure 2-9. It can be observed from this figure that the magnitude frequency response of a digital anti-resonator is a mirror image of the magnitude frequency response of digital resonator with the same center frequency and bandwidth.

### 2.4.8.3 Cascade filter bank

In a cascade filter bank the second order digital resonators and the digital anti-resonators are connected in series. The transfer function for 'P' resonators and 'Z' anti-resonators connected in series is given by:

$$T(z) = \prod_{i=1}^P \frac{A_i}{1 - B_i z^{-1} - C_i z^{-2}} \prod_{j=1}^Z A'_j + B'_j z^{-1} + C'_j z^{-2}$$

where the values of the digital resonator coefficients  $A_i$ ,  $B_i$  and  $C_i$  are obtained from the  $i$ th formant frequency and bandwidth parameters and the values of the digital anti-resonator coefficients  $A'_j$ ,  $B'_j$  and  $C'_j$  are obtained from the  $j$ th anti-formant frequency and bandwidth parameters.

#### 2.4.8.4 Parallel filter bank

In a parallel filter bank the second order digital resonators and the digital anti-resonators are connected in parallel. Each resonator or anti-resonator is preceded by a multiplier whose gain is specified by the formant or anti-formant amplitude control parameter. The transfer function for 'P' resonators and 'Z' anti-resonators connected in parallel is given by:

$$T(z) = \sum_{i=1}^P \frac{a_i A_i}{1 - B_i z^{-1} - C_i z^{-2}} \sum_{j=1}^Z a'_j (A'_j + B'_j z^{-1} + C'_j z^{-2})$$

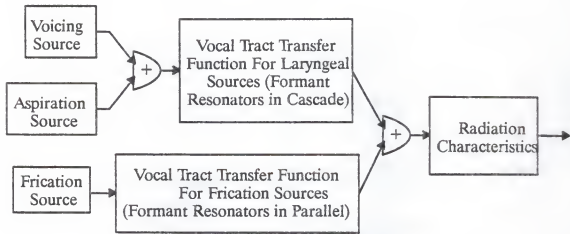
The parameters " $a_i$ " and " $a'_j$ " are the values of the  $i$ th and the  $j$ th formant and anti-formant amplitude parameters associated with the  $i$ th and  $j$ th resonator and anti-resonator, respectively.

#### 2.4.8.5 Cascade versus parallel filter bank

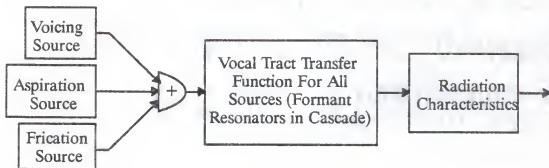
The historic development of the formant synthesizer (see Appendix A) shows that there have been opposing viewpoints with respect to the cascade and the parallel connections of the resonators for the synthesis of voiced sounds. Among the recent and commonly used implementations of the formant synthesizers is Klatt's cascade/parallel formant synthesizer uses the cascade filter bank to synthesize voiced sounds, whereas the Holmes' all-parallel formant synthesizer uses the parallel filter

bank with independent gain controllers to synthesize the various sounds. Klatt has argued that with the cascade configuration the relative amplitudes of the formant peaks for vowels come out just right [Fant, 1956] without the need of individual amplitude controls. Also, the cascade configuration is a more accurate model of the vocal-tract transfer function during the production of vowels [Flanagan, 1957]. Holmes has argued that, when properly implemented, the parallel configuration is in fact superior in all significant respects for synthesis of both vowels and consonants. He has shown that the actual vocal-tract transfer function departs from the theoretical transfer function obtained under the assumption of plane wave propagation of sounds in the vocal tract [Fant, 1960]. He argues that the cascade model can give a good approximation of the transfer function only up to 3 KHz. Also, the critical bands in the frequency region above 3 KHz are of the order of 500 Hz wide, so the spectral fine details of the speech signal above 3 KHz are not perceptually important. Only the levels of the magnitude response are perceptually important, and there is no obvious reason for the "pole" frequencies of the filter bank to be the same as the resonance frequencies of the vocal tract in this part of the speech spectrum.

Klatt's cascade/parallel formant synthesizer is a compromise of both the views. This synthesizer can be used in either the cascade/parallel synthesizer configuration or in a special purpose all-parallel synthesizer configuration. The block diagrams of these two synthesizer configurations are given in Figure 2-10. In the cascade/parallel synthesizer configuration the cascade filter bank is used to synthesize voiced, aspirated, mixed and nasal sounds and the parallel filter bank is used to synthesize plosives and fricatives. In the all-parallel synthesizer configuration, as in the Holmes' all-parallel formant synthesizer, the parallel filter bank is used to synthesize all types of sounds. The user must decide, before starting the synthesis process, which of the two configurations of the synthesizer should be used for synthesis.



[a]



[b]

Figure 2-10: Two configurations of Klatt's cascade/parallel formant synthesizer  
 a) Cascade/parallel synthesizer configuration  
 b) All-parallel synthesizer configuration



#### 2.4.8.6 Filter Banks in the flexible formant synthesizer

In the flexible formant synthesizer, while specifying the parameters, the resonators, the anti-resonator and the multipliers (by-pass paths) are uniformly treated as second order filters. For the  $i^{\text{th}}$  filter, the center frequency is specified by the parameter “fi,” the filter bandwidth by the parameter “bi” and the filter gain ( $a_iA_i$ ) by the amplitude control parameter “ai.” For a multiplier, the center frequency and bandwidth parameters are specified equal to zero, and hence, the filter coefficients ‘B’ and ‘C’ are equal to zero and ‘A’ is equal to 1. The multiplier’s gain is specified by the amplitude control parameter “ai” associated with the filter simulating the multiplier. The filter bank configuration is specified through “filter specifications.” A filter specification for each filter has the following information:

- 1) the filter-number assigned to the filter,
- 2) the type of the filter: a resonator, anti-resonator or a multiplier,
- 3) the filter bank(s) it belongs to: cascade, parallel or both,
- 4) the type of excitation source,
- 5) the scale factor for specifying the “initial phase” of the filter output, and
- 6) the center frequency of the filter.

The filter-number assigned to a filter is the same as the parameter symbol name used to specify the center frequency of that filter. Thus, the parameter, “fi,” for specifying the center frequency of the  $i^{\text{th}}$  filter is also used to specify all the information about that filter. The user assigns a filter-number and specifies all the other relevant information for each formant, anti-formant and by-pass path through filter specifications. The user is not restricted to assigning a particular filter and filter-number to a particular formant, anti-formant or multiplier. Thus, the filters (resonators, anti-resonators and multipliers) in the filter bank are not labelled and pre-assigned to generate specific formants and anti-formants; instead each filter is

labelled by the filter-number in its filter specification. For example, the filter-number 1 does not necessarily correspond to the first formant generator. A filter is configured as a resonator, anti-resonator or multiplier based upon the filter type specified in its filter specification. A resonator is specified to generate a formant and an anti-resonator is specified to generate an anti-formant. A multiplier is specified as a by-pass path in the parallel filter bank and as a gain controller in the cascade filter bank. A filter may be specified to belong to either the cascade filter bank only, the parallel filter bank only or to both the cascade and parallel filter banks. If a filter is specified to belong to both the filter banks then two identical filters with the same filter coefficients are created in each filter bank. For reasons discussed later, an anti-resonator cannot be specified for the parallel filter bank.

The cascade filter bank has a fixed excitation source. Each filter in the parallel filter bank may have a separate excitation source. In both Klatt's and Holmes' synthesizers, a filter in the parallel filter bank has an excitation source based upon the formant generated by that filter. In the flexible formant synthesizer, the user can select any one of the ten available excitation sources (obtained by various combinations and modifications of the voicing source, aspiration noise source and the friction noise source) as an input to any filter in the parallel bank. The scale factor ( $\pm 1$ ) specifies the "initial phase" ( $0$  or  $\pi$ ) of the filter output. The values of the scale factors are restricted to  $+1$  and  $-1$  in order to restrict the values of "initial phase" to  $0$  and  $\pi$  radians, respectively. The center frequency of the filter (a resonator or an anti-resonator) can either be constant or variable during the synthesis of an utterance. The procedure to assign the filter-numbers to the formants, anti-formants and multipliers, and to specify all the information required by the filter specifications is described in the "Formant Synthesizer Users Manual."

At start-up (initiation of synthesis), the user specifies all the information about the formants, anti-formants and by-pass paths required for synthesizing a particular

token through a list of filter specifications. The number of filters in the cascade filter bank and/or the parallel filter bank are determined from this list. The resonators and anti-resonators in the filter bank(s) in the flexible formant synthesizer are labelled by their filter-numbers (i.e., by the parameters for specifying the center frequencies of the filters). The cascade and/or parallel filter banks are configured accordingly. The assignments of formants and anti-formants to the filters in the filter bank(s) remains fixed until the synthesis is completed. When the center frequency of a filter is specified as a constant, its value specified in the filter specification at the start-up is used during the synthesis. When the center frequency parameter is variable, the value of the center frequency parameter is updated at the beginning of each frame. The filter coefficients are calculated from the formant or anti-formant frequency and the bandwidth parameters. The filter coefficients are assigned to the filter(s) in the filter bank(s) generating that formant or anti-formant.

The default configuration for the filter banks in the flexible formant synthesizer is nearly the same as that of the filter banks in the Klatt's cascade/parallel formant synthesizer. The default configuration of the cascade and the parallel filter banks is shown in Figure 2-11.

#### 2.4.8.7 Default configuration of the cascade filter bank

The default configuration of the cascade filter bank in the flexible formant synthesizer is the same as the cascade filter bank in Klatt's cascade/parallel formant synthesizer. This configuration is designed to synthesize voiced speech signals with bandwidths up to 5 KHz. Klatt and Holmes have argued, based on the acoustic theory of speech production [Fant, 1960], that for the non-nasalized voiced sounds and aspirated sounds (both with the source at the glottis and bandwidth up to 5 KHz) the vocal-tract transfer function can be adequately represented by only "poles" (resonators) and no "zeros" (anti-resonators) [Klatt, 1980; Holmes, 1983]. Five

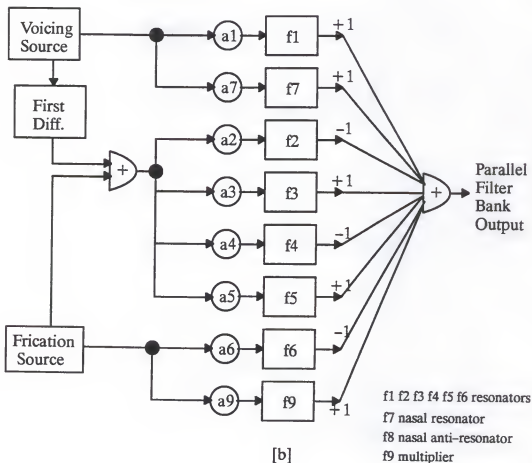
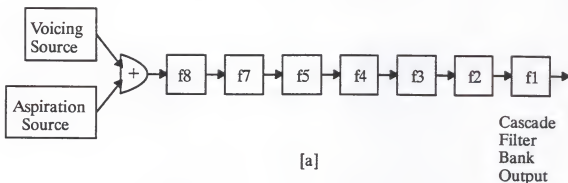


Figure 2-11: Default configurations of the filter banks in the flexible formant synthesizer  
a) Cascade filter bank  
b) Parallel filter bank

formants (resonators or “poles”) for male voice and four formants for female voice are normally adequate to represent the non-nasalized voiced sounds. For nasal sounds (nasal murmur and nasalized vowels), the transfer function should include an additional resonator and anti-resonator pair to represent the nasal formant and anti-formant pair near the first formant frequency. The center frequency (270 Hz) and bandwidth (250 Hz) are kept equal for both the nasal resonator and the nasal anti-resonator in order to cancel each others effects when synthesizing non-nasalized voiced sounds. For synthesizing nasals, the center frequency of the nasal anti-resonator is shifted to the average value of the nasal resonator’s constant center frequency and the constant/variable first formant frequency.

The overall frequency response of the cascade filter bank is the product of the frequency responses of the individual filters (resonators and anti-resonators). If the center frequencies of the filters (both the resonators and the anti-resonators) are spaced reasonably for apart and the bandwidths are not too large, it is possible to obtain peaks (formants) and notches (anti-formants) of desired widths and at the desired frequency locations in the overall magnitude frequency response of the cascade filter bank. The amplitude of the formants (peaks) and anti-formants (notches) in the magnitude frequency response depend upon the center frequency and bandwidth of the filters in the cascade filter bank and cannot be independently controlled.

#### 2.4.8.8 Default configuration of the parallel filter bank

The default configuration of the parallel filter bank in the flexible formant synthesizer is similar to that of the parallel filter bank in Klatt’s cascade/parallel formant synthesizer. The parallel filter bank can be used to synthesize sounds whose 5 KHz spectrum contains both the formants and anti-formants. For such sounds, the vocal tract transfer function is represented by both “zeros” and “poles.” Although,

a cascade filter bank may be used to create "poles" and "zeros" in the transfer function, the amplitudes of the formants and anti-formants in the magnitude frequency response are not properly determined for these sounds. This is due to the fact that the amplitude of the formants and anti-formants cannot be controlled when the cascade filter bank is used. Klatt (1980) and Holmes (1983) have argued that for non-nasalized sounds up to 5 KHz five formants (resonators) are sufficient to represent most speech sounds. The perceptual effect of anti-formants in the speech sound spectrum is not significant because the masking effect of the energy in the adjacent formant peaks limit the detectability of a spectral notch (anti-formant). Therefore, the parallel filter bank in Klatt's cascade/parallel formant synthesizer and Holmes' all-parallel formant synthesizer do not employ anti-resonators to create anti-formants in the magnitude frequency response. The default configuration of the parallel filter bank in the flexible formant synthesizer does not contain anti-resonators.

In the cascade/parallel synthesizer configuration, the parallel filter bank is used to synthesize only fricatives. In the all-parallel synthesizer configuration the parallel filter bank is used to synthesize all types of sounds. Klatt has mentioned that for synthesizing non-nasalized voiced sounds, five resonators are adequate. However, the parallel filter bank in Klatt's synthesizer employs only four resonators to synthesize these sounds. The parallel filter bank in the flexible formant synthesizer employs five resonators for synthesizing these sounds, as observed in Figure 2-4b. For these sounds, the "poles" in the vocal-tract transfer function can be regarded as representing true formant resonances. During the synthesis of aspirated sounds and fricatives, five resonators (corresponding to the second to sixth formants) are used. For these sounds, the "poles" with large bandwidths in the vocal-tract transfer function can be regarded as representing the broadband levels in the spectra of these sounds. The amplitude of each formant/spectral level is independently controlled by

the amplitude control parameter associated with the resonator generating that formant/spectral level. A sixth resonator (formant) is added to the parallel filter bank specifically for simulating the high-frequency spectral levels. A by-pass path with gain (amplitude) controller (multiplier) is included in the parallel filter bank to simulate the spectra of the sounds that contain no prominent formants (peaks) or anti-formants (notches). For the nasal sounds, the nasal formants near the first formant frequency is represented by an additional resonator with a fixed center frequency (270 Hz) and bandwidth (250 Hz). The value of the amplitude control parameter of this resonator is normally kept low and is set to a high value during the synthesis of nasals. The technique to create a nasal anti-formant is described later.

The overall frequency response of the parallel filter bank is the sum of the frequency response of the individual filters. The overall frequency response near a formant peak depends largely upon the frequency response of the resonators generating that formant and also upon the skirt responses of all the other resonators in the filter bank. A skirt response of a filter can be described as the frequency response of the filter at frequencies much higher and lower than its center frequency. To reduce the number of control parameters required, only the amplitude control parameter associated with each resonator should control the amplitude of the formant/spectral-level generated by that resonator. To achieve this, the amplitude of the skirt response of all the resonators should be highly attenuated by associating each resonator with additional formant-shaping filters. Holmes (1983) has discussed a set of ideal and practical conditions that should be satisfied by these formant-shaping filters. Klatt (1980) and Holmes (1983) have shown that, in the high frequency range of the overall frequency response (above the normal range of the first formant frequency), an efficient means of cancelling (attenuating) the skirt responses of the resonators is to add the output from the resonators with adjacent center frequencies in opposite polarities. This method also prevents creation of



anti-formants in between the two formants (to be explained later). However, at low frequencies (below the first formant frequency), it is difficult to achieve both the cancellation of skirt responses of the first and the second formant generators (resonators) for voiced sounds, and also achieve a considerable attenuation of the low frequency skirt response of the second and higher formant generators for voiceless consonants. Both Klatt's and Holmes' synthesizers have employed a first order FIR filter as a frequency-shaping filter. A first order FIR filter is connected in series with all the resonators in the filter bank except for the first formant generator (resonator). With this method, it is possible to achieve considerable attenuation below the second formant frequency (when the first formant is not used), so that the voiceless consonant spectra can be adequately represented. Holmes (1983) provides a phase correction circuit in series with the first formant generator so that the skirt responses of the first and the second formant generators cancel properly during the synthesis of voiced sounds.

Each resonator in the parallel filter bank is a linear system. Also, these resonators have a common excitation source. Therefore, a single first order FIR filter is placed in series with the common excitation source instead of separate first order FIR filters at the output of each resonator. In the flexible formant synthesizer, we employ a similar strategy. The parameter "ph\_filt" specifies the coefficient of the first order FIR filter used for differentiating the excitation source. The assignment of the excitation source for the resonators in the parallel filter bank is as follows: 1) the excitation source for the first formant and nasal generator is the voicing source alone, 2) the excitation source for the second, third, fourth and fifth formant generators is a sum of the differentiated glottal source and the noise source, and 3) the excitation source for the sixth formant generator and the by-pass path is the noise source alone. The magnitude frequency response of the parallel filter bank when simulating uniform tube and the vowels /t/ and /a/ with/without the first order differentiator in series with



the resonators is shown in Figure 2-12. Since the output from the resonators in the parallel filter bank with the adjacent center frequencies are added with alternate polarities, anti-formants are not observed in between the formants. However, the magnitude frequency response is not properly determined at the low frequencies when a differentiator is not used.

#### 2.4.8.9 Creating zeros (anti-formants) in the parallel filter bank

The anti-resonators cannot create anti-formants ("zeros") in the magnitude frequency response of a parallel filter bank. The reason being that anti-resonators have a high-amplitude high-frequency skirt response. If the output scale-factors of two resonators have the same polarities ("initial phase" values have a difference of zero radians), the transfer function of these two resonators in the parallel configuration has a "zero" in between the two "poles" at the center frequencies of the resonators [Flanagan, 1957]. If the output scale factors of the two resonators have the opposite polarities ("initial phase" values have a difference of  $\pi$  radians), the transfer function of these two resonators in the parallel configuration is similar to that of the transfer function of these two resonators in the cascade configuration [Klatt, 1980]. The magnitude frequency response in the first case has an anti-formant in between the two center frequencies of the two resonators. The magnitude frequency response in the latter case has peaks at the center frequencies of the two resonators and no anti-formant in between. The magnitude frequency response of the two resonators in the cascade configuration and the above mentioned two cases of the parallel configuration are shown in Figure 2-13. The "initial phase" value of the filter's output can be set equal to 0 or  $\pi$  radians by specifying the value of the scale factor for the filter output to be  $\pm 1$  in the filter specification. The user can create an anti-formant (notch) in between two formants (peaks) in the magnitude frequency

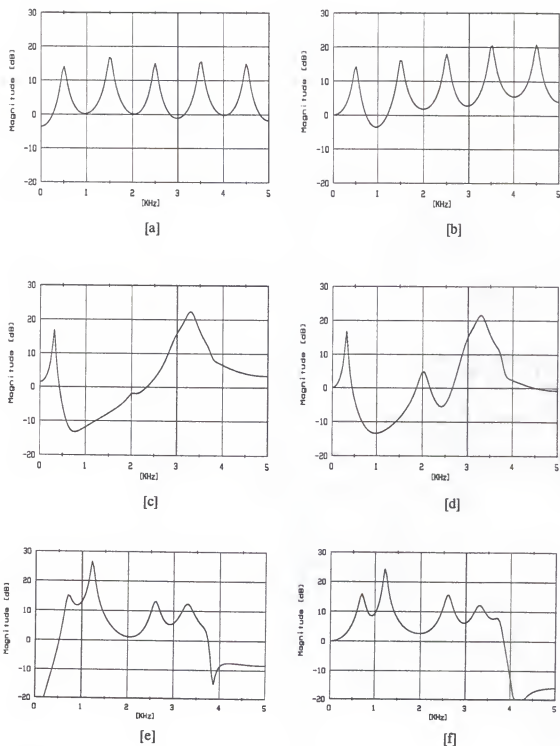
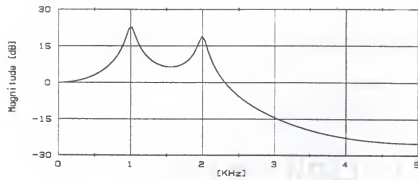
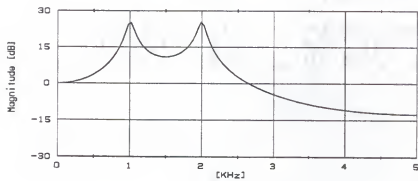


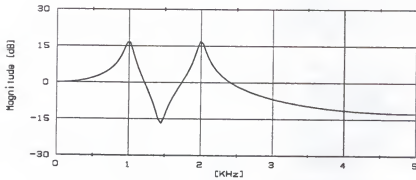
Figure 2-12: Magnitude frequency response of parallel filter bank  
 (without differentiator (left) and with differentiator (right))  
 a) and b) Uniform tube  
 c) and d) Vowel /i/ (using typical values of formant parameters)  
 e) and f) Vowel /a/ (using typical values of formant parameters)



[a]



[b]



[c]

Figure 2-13: Magnitude frequency response of two resonators in the  
 a) Cascade configuration  
 b) Parallel configuration with opposite polarities of output  
 c) Parallel configuration with same polarities of output

response of the parallel filter bank by specifying the same values of the scale factors for the two resonators generating these two formants.

#### 2.4.8.10 Simulation of a cascade filter bank with a parallel filter bank

In both the cascade/parallel synthesizer configuration and the all-parallel synthesizer configuration, during the synthesis of voiced sounds, the magnitude frequency response of the parallel filter bank should match the magnitude frequency response of the cascade filter bank. This can be achieved by 1) specifying each resonator in the cascade filter bank to also be a resonator of the parallel filter bank and 2) controlling the amplitude control parameters of the resonators in the parallel filter bank in order to match the amplitudes of the corresponding formants (peaks) in the magnitude frequency response of the cascade filter bank. The transfer function of the cascade filter bank can be expanded in partial fractions, in which, each term represents the frequency response of a single resonator. When the gain of each resonator ( $a_i A_i$  for the  $i$ th resonator) in the parallel filter bank is set equal to the partial fraction coefficient, the transfer function of the parallel filter bank is equivalent to the transfer function of the cascade filter bank. Holmes has shown that even a slight change in the partial fraction coefficient of any resonator can cause deleterious effects on the overall magnitude frequency response of the parallel filter bank. Hence this method for matching the magnitude frequency response of the two filter banks is not used.

Klatt (1980) has described a procedure to obtain a match at the formant peaks and also in the low frequency region of the magnitude frequency response of the cascade and the parallel filter banks. The first step in this procedure involves adding the scale factor to the amplitude control parameter of each of the resonators in the parallel filter bank. The values of the scale factors are adjusted such that the magnitude frequency response of the parallel filter bank is the nearly the same as that

of the uniform tube. The magnitude frequency response of the uniform tube has all the resonance peaks at the same amplitude and bandwidth, and occurring at equal intervals along the frequency axis. In the next step, a set of rules, described in Klatt (1980), are used to modify the values of these scale factors. These modifications are proportional to the relative shifts of the center frequencies and bandwidths of the resonators from the resonance frequencies and bandwidths of the uniform tube.

Let the five resonators in the cascade and parallel filter banks have center frequencies at 500, 1500, 2500, 3500 and 4500 Hz, bandwidths equal to 100 Hz and amplitude control parameters equal to 60 dB. The magnitude frequency response of the cascade filter bank is similar to that of the uniform tube as seen in Figure 2-14a. The magnitude frequency response of the parallel filter bank is shown in Figure 2-14b. The formants (peaks) are produced at the specified center frequencies and have equal bandwidths. The formant amplitudes increase with the increasing center frequency, since, the "Q factor" (ratio of center frequency to bandwidth) of the resonators is increasing. Klatt has given a set of scale factors, which, when added to the amplitude control parameters of the resonators in the parallel filter bank, should generate the magnitude frequency response with equal amplitude of the formants. This magnitude frequency response is shown in Figure 2-14c. Comparing the magnitude responses in Figure 2-14a and c, we observe that the formant amplitudes are still not properly determined. However, if the first order FIR filter is not used in series with the resonators in the parallel filter bank, the amplitude of the formants are approximately equal as observed in Figure 2-14d. Therefore, we decided to modify Klatt's scale factors in order to obtain equal amplitude formants when the first order FIR filter is used in the parallel filter bank. The value of each of the new scale factors is set equal to the difference (in dB) between the values of the corresponding peaks in the magnitude frequency response shown in Figure 2-14a and Figure 2-14b. The magnitude frequency response of the parallel filter bank when the new scale factors

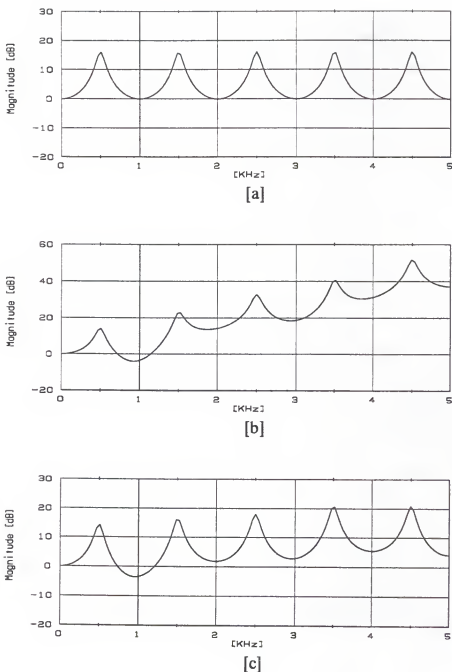


Figure 2-14: Simulation of magnitude frequency response of a uniform tube by the parallel filter bank

- a) Cascade filter bank simulating the uniform tube
- b) No scale factors
- c) Using Klatt's scale factors and rules
- d) Same as (c) without the first order FIR filter
- e) Using the new scale factors and Klatt's rules
- f) Using the new procedure

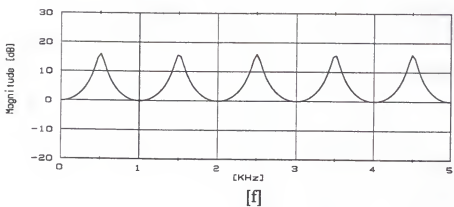
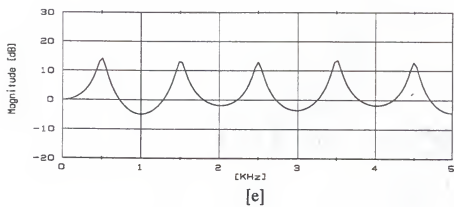
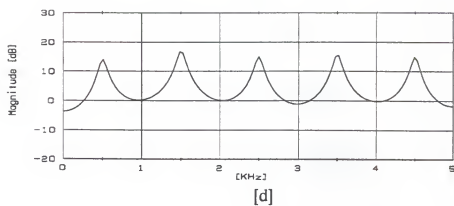


Figure 2-14: Continued

are added to the amplitude control parameters of the resonators is shown in Figure 2-14e. We observe that the new scale factors give a better match between the magnitude frequency response of the parallel filter bank using the first order FIR filter and the magnitude frequency response of the uniform tube. The magnitude frequency response of both the cascade filter bank and the parallel filter bank (using Klatt's scale factors and rules) with the typical values of the formant frequencies and bandwidths for the vowel /t/ are shown in Figure 2-15 and for the vowel /a/ in Figure 2-16. We observe that the magnitude frequency response of the parallel filter bank obtained by using the modified scale factors show a better match with the magnitude frequency response of the cascade filter bank. The magnitude frequency response of the parallel filter bank using the new set of scale factors when simulating the uniform tube, and vowels /t/ and /a/ with/without the first order FIR filter in series with the resonators are shown in Figure 2-17. We can observe that the magnitude frequency response is not properly determined at the low frequencies when the differentiator is not used. In the flexible formant synthesizer, we have included both the Klatt's scale factors and the new scale factors. The user can select either set of scale factors by specifying an appropriate option at start-up as explained in the "Formant Synthesizer Users Manual." This method for automatically adjusting the amplitude control parameters of the resonators in the parallel filter bank is fast. But, this procedure has several limitations: 1) it is inadequate when the center frequencies have large shifts, 2) the relative shifts in the bandwidths are not considered in Klatt's rules, and 3) it can be used only when exactly five resonators are specified for a signal bandwidth up to 5 KHz.

We have developed another procedure for matching the magnitude frequency response of the cascade filter bank by the parallel filter bank. The amplitude of each formant in the magnitude frequency response of the cascade filter bank depends upon the center frequencies and bandwidths of all other resonators. The amplitude control



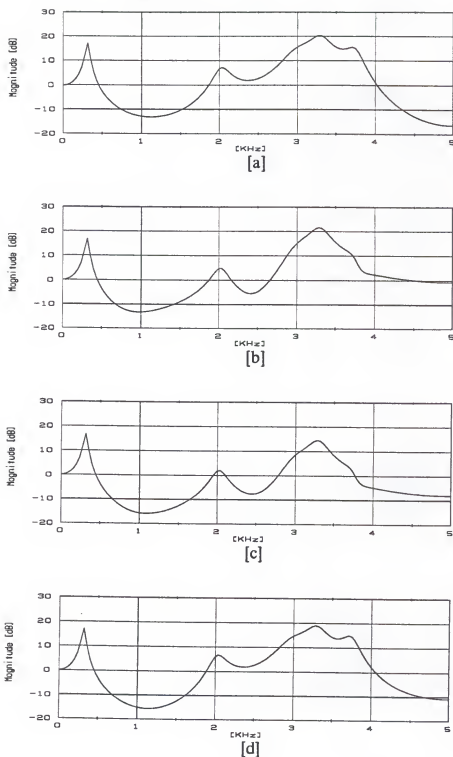


Figure 2-15: Simulation of magnitude frequency response of vowel /i/ by the parallel filter bank

a) Cascade filter bank

b) Using Klatt's scale factors and rules with parallel filter bank

c) Using new scale factors and Klatt's rules with parallel filter bank

d) Using new procedure with parallel filter bank

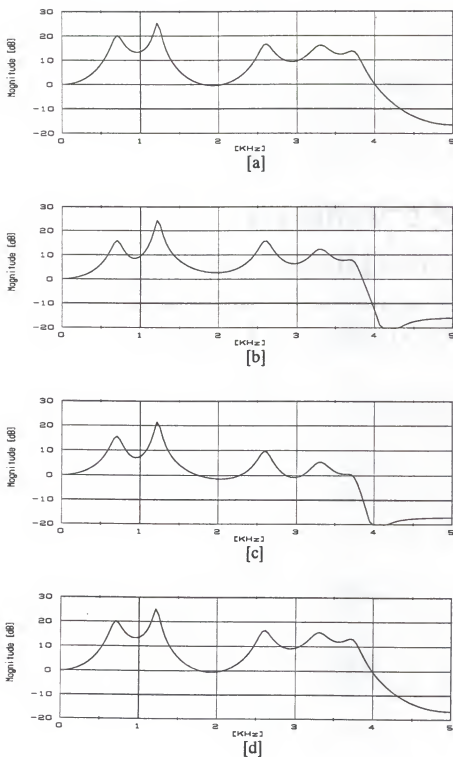


Figure 2-16: Simulation of magnitude frequency response of vowel /a/ by the parallel filter bank

- Cascade filter bank
- Using Klatt's scale factors and rules with parallel filter bank
- Using new scale factors and Klatt's rules with parallel filter bank
- Using new procedure with parallel filter bank

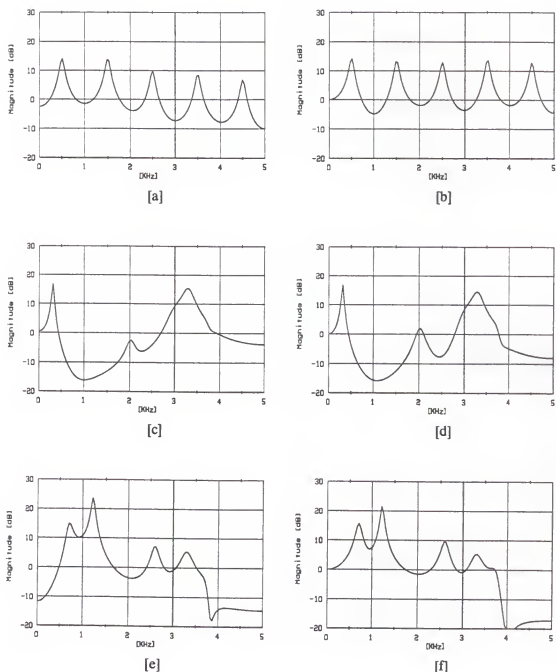


Figure 2-17: Magnitude frequency response of parallel filter bank with the new scale factors for the amplitude control parameters (without differentiator (left) and with differentiator (right))  
 a) and b) Uniform tube  
 c) and d) Vowel /t/ (using typical values of formant parameters)  
 e) and f) Vowel /a/ (using typical values of formant parameters)

parameters have no effect on the magnitude frequency response of the cascade filter bank. The exact magnitude frequency response of the cascade filter bank can be calculated using the transfer function of the cascade filter bank. In the new procedure, first the amplitude of the magnitude frequency response of each resonator at the center (formant) frequency of each resonator in the cascade filter bank is calculated. Then the amplitude (in dB) of each formant in the magnitude frequency response of a cascade filter bank is calculated by summing the amplitude of the magnitude frequency response (in dB) of all the resonators in the cascade filter bank at the formant frequency.

In the flexible formant synthesizer, the parallel filter bank consists of resonators only. The amplitude of a formant in the magnitude frequency response of the parallel filter bank is determined by the "Q factor" of that formant generator (resonator), when the skirt responses of the other resonators are negligible at that formant frequency. In the new procedure, the value of the scale factor for the amplitude control parameter associated with the resonator common to both the filter banks is found such that the sum of the "Q factor" of that formant resonator and the scale factor for the amplitude control parameter of that resonator is equal to the amplitude of the corresponding formant in the magnitude frequency response of the cascade filter bank. When the values of all the amplitude control parameters are equal to zero dB and the scale factors (in dB) are added to the amplitude control parameters (in dB) of the resonators in the parallel filter bank, the amplitude of the formants in the magnitude frequency response of the parallel filter bank match with the amplitude of the corresponding formants in the magnitude frequency response of the cascade filter bank. The second step for further modification of scale factors is not required.

For those resonators in the parallel filter bank that do not have a duplicate in the cascade filter bank (this case does not occur in Klatt's cascade/parallel synthesizer configuration), the amplitude at the "poles" in the transfer function is determined by

the resonator's "Q factor," scale factor and the value of the amplitude control parameter. In order that the amplitude of the "pole" is determined only by the amplitude control parameter, the scale factor is computed to be equal to the negative of the value of the "Q factor." For voiced sounds, the amplitude control parameters of the resonators in the parallel filter bank should be kept equal to 0 dB. The scale factor for the "initial phase" of the output of the resonators should be assigned +1 and -1 alternately, to avoid creation of the anti-formants in the magnitude frequency response. The Figure 2-14, Figure 2-15 and Figure 2-16 show the matching of the magnitude response of the cascade and parallel filter banks for the uniform tube, and vowels /i/ and /a/ with this new procedure. The magnitude frequency response obtained by the new procedure matches exactly at the formants and roughly in between.

The new procedure requires more computation than the procedure proposed by Klatt. But, this method is more accurate and is not limited to only five resonators with a signal bandwidth of 5 KHz. Also, the values of the scale factors for the amplitude control parameters are dynamically adjusted at the beginning of each frame, and therefore, a good match between the magnitude frequency response of the cascade and parallel filter banks is achieved for all frames.

### 2.5 Radiation Load

The volume-velocity radiated at the lips and nose creates a sound pressure wave. It has been observed that the frequency distribution of the sound pressure,  $P(f)$ , radiated from the head, at sufficiently far distance can be approximated by a highpass filtering of the volume-velocity at the lips,  $U(f)$  [Fant, 1960]. The radiation load,  $R(f)$ , can thus be represented by a first order FIR filter placed at the output of the filter bank(s) simulating the vocal-tract transfer function,  $T(f)$ . Since these filter banks are

linear systems, the same radiation-load filter can also be placed in series with the noise source model and the glottal source model. In this case, the first order FIR and IIR filters in cascade with the noise source model effectively cancel each other, and hence, both the filters can be removed from the synthesizer architecture. A series combination of the glottal source model and the first order FIR filter produces differentiated glottal source pulses. This is how the radiation load is represented in Klatt's cascade/parallel synthesizer.

In the flexible formant synthesizer, several types of glottal source models and the noise source models are incorporated. Therefore, the user should have the flexibility of being able to select the type of filter to be used to modify the glottal source, noise source or the radiation load. The user should also have the flexibility to modify the spectra of various signals at intermediate stages in the synthesizer. Therefore, we have placed several FOSs in the synthesizer architecture in order to modify the spectra of various intermediate signals. As seen earlier, with a FOS, the user can specify a first order FIR filter, a first order IIR filter or may choose not to use any such filter. Therefore, using FOS, the user can specify any desired type of filter and a variable filter coefficient to modify the glottal source spectrum, noise source spectrum, the radiation load characteristics and spectra of various intermediate signals. The parameter "g\_filt" specifies the filter coefficient for the FOS in series with the glottal source model, the parameter "n\_filt" specifies the filter coefficient for the FOS in series with the noise source model and the parameter "o\_filt" specifies the filter coefficient for the FOS at the output of the filter banks. The synthesizer block diagram shows, several other FOS systems placed in various sections of the flexible formant synthesizer architecture. The default configuration of the synthesizer, which is Klatt's cascade/parallel formant synthesizer, has a FIR filter in series with the glottal source model and no filters (FIR or IIR) in series with the noise source model at the output of the filter banks.

## 2.6 Synthesis Strategies

In the following section we outline some of the basic procedures used to synthesize speech using the default configuration of the flexible formant synthesizer. One should realize that it is not possible for us, in this study, to outline the general synthesis strategies, such as those used for the synthesis of English syllables, words, etc., in the rule-based speech synthesis systems. The strategies outlined here are simple steps to be followed to synthesize simple utterances. The exact synthesizer parameters and synthesis procedures depend upon the speech tokens to be synthesized and the context in which such speech tokens will be used. For example, if high-quality synthetic speech is desired then the synthesizer parameters should be accurately specified and the synthesis procedure should be carefully controlled. In the following sub-sections, we describe how the excitation source (glottal and noise) are generated and the filter banks (cascade and/or parallel) are configured to synthesize voiced, unvoiced and mixed sounds.

### 2.6.1 Voiced Sounds

A glottal source model is used as an excitation source for synthesizing vowels (/i/, /ɪ/, /e/, /æ/, /a/, /ɜ/, /ʌ/, /ɔ/, /u/, /ʊ/ and /o/), semi-vowels (/w/, /l/, /r/ and /j/), diphthongs (/aɪ/, /ɔɪ/, /aʊ/, /eɪ/, /oʊ/ and /ju/) and nasals (/m/, /n/ and /ŋ/). The glottal source model generates the glottal source pulses that simulate the volume-velocity pulses of air produced by the quasi-periodic vibrations of the vocal folds as the air flows from the lungs to the pharynx. The shape of the glottal pulses is controlled by the parameters of the glottal source model. The shape of the glottal pulses determine the vocal characteristics of synthesized speech, such as breathy, normal, etc. A detailed explanation of how to control the values of the parameters of the glottal source models (discussed in Appendix B) to obtain various glottal pulse shapes is

beyond the scope of this section. The readers are referred to the references listed with the description of the glottal source models in Appendix B.

In the flexible formant synthesizer, the rate of generation of the glottal pulses (fundamental frequency) is specified by the parameter "f0." The higher the value of the parameter "f0," the higher is the value of perceived "pitch." The energy/power in each of the glottal pulses is specified by the parameter "av." The higher the value of the parameter "av," the higher is the perceived "stress." The duration of a voiced sound is determined by the number of consecutive frames for which both the "av" and "f0" parameters are specified to be nonzero. The gain and the pitch contours for an utterance represent the variation in the values of the "av" and "f0" parameters during the utterance. It is the pitch contour, gain contour and the duration of each sound in an utterance that determines the "intonation" and "stress" patterns of an utterance.

According to the acoustic theory of speech production, the non-nasalized voiced sounds can be represented by an "all-pole" vocal-tract transfer function. Different voiced sounds are produced by different vocal tract configurations and movements of articulators, and therefore, have different formant parameters (frequencies, bandwidths and amplitudes). In the American English language, most voiced sounds can be adequately represented by the lower three formant frequencies and bandwidths. Over the duration of an utterance, there is a wide range of variation in the values of the formant frequencies and amplitudes but only slight variation in the values of formant bandwidths. Also, there is variation in the inherent duration of each voiced sound. Typical values of formant frequencies and bandwidths for the voiced sounds have been established by analysis of several speech tokens for males and females [Peterson and Barney, (1952); Klatt, 1980; Childers and Wu 1990]. Also, a table of minimum and inherent durations of the voiced sounds is given in Klatt (1987). Several researchers have shown that a fairly good copy of an all voiced utterance can



be obtained by replicating the formant tracks (formant frequencies and bandwidths) of the first three formants, and the pitch and gain contours.

For synthesizing sustained phonations, such as sustained vowels /a/, /i/, etc., the "av," "f0," "f1," "f2," "f3" "b1," "b2" and "b3" parameters should be specified as constant. For synthesizing an all-voiced multiple sounds utterance the values of the above parameters are variable and are specified through parameter tracks. Default values of the fourth and fifth formant frequencies and bandwidths can be used during the synthesis. If the parallel filter bank is used to synthesize an all-voiced utterance, the amplitude control parameters of all the resonators in the parallel filter bank should be set to zero, since they are automatically determined (to make the magnitude frequency response of the parallel filter bank equivalent to that of the cascade filter bank), unless variations in the "normal" amplitude of the formants is desired. The frame size may be kept constant as specified by the parameter "frame\_size," or can be equal to the pitch-period, if the "PITCH\_SYNC" flag is set to one. For sustained phonations, the total number of frames to be synthesized should be specified by the parameter "tot\_frames." For the multiple sound utterance, the total number of frames to be synthesized is equal to the length of the parameter tracks. The values in the parameter tracks are used to update the values of these parameters at the beginning of each frame.

When the cascade filter bank is used to synthesize an utterance with the nasal sounds, a parameter track for the center frequency of the nasal anti-resonator should be specified. This parameter track should have a constant value of 250 Hz (default value) except for the duration of a nasal murmur or a nasalized vowel. For the duration of the nasal murmur, the center frequency of the nasal anti-resonator should be shifted to 450 Hz [Klatt, 1980]. For the duration of nasalized vowel, a "pole-zero-pole" combination is created using nasal resonator, nasal anti-resonator and first formant generator (resonator). The first formant frequency should be

increased by 100 Hz and the center frequency of the nasal anti-resonator should be equal to the average value of the the center frequency of the nasal resonator (fixed at 250 Hz) and the first formant frequency [Klatt, 1980]. When using the parallel filter bank and Klatt's rules (with either Klatt's scale factors or the new scale factors) for synthesizing the nasal murmur and nasalized voiced sounds, the amplitude control parameter of the nasal resonator has to be set to about 60 dB for the duration of the nasal sound and zero dB otherwise. When using the new procedure the amplitude control parameter of the nasal resonator should be kept zero since the amplitude of the nasal formant is set equal to the amplitude of the nasal resonator in the cascade filter bank. For both the procedures, the amplitude control parameter of the other resonators should be kept zero, unless variations in the "normal" amplitude of the formants is desired.

In the cascade/parallel synthesizer configuration only the cascade filter bank is used for synthesis of a sustained vowel and an all voiced sentence. Therefore, we have created an additional synthesizer configuration "all-cascade synthesizer configuration," in which, only the cascade filter bank is used for synthesis. The type of synthesizer configuration (filter bank(s)) used for synthesizing an utterance is specified by the parameter "arch\_typ." An example of a synthesized speech signal, its spectrum and the spectrum of the natural speech signal for sustained vowel /ɪ/ is shown in Figure 2-18. Figure 2-19 shows the spectrogram of the synthesized and natural utterance of the sentence "We were away a year ago." A visual comparison of the spectra in Figure 2-18 and spectrograms in Figure 2-19 show a good match between the synthesized and natural speech utterances. The formant frequencies and bandwidths were estimated from the LPC analysis. Since the LPC analysis tend to underestimate formant bandwidths, the formant bandwidths appear to be larger in the spectrogram of natural utterance as compared to the formant bandwidths in the spectrogram of synthetic utterance in Figure 2-19.

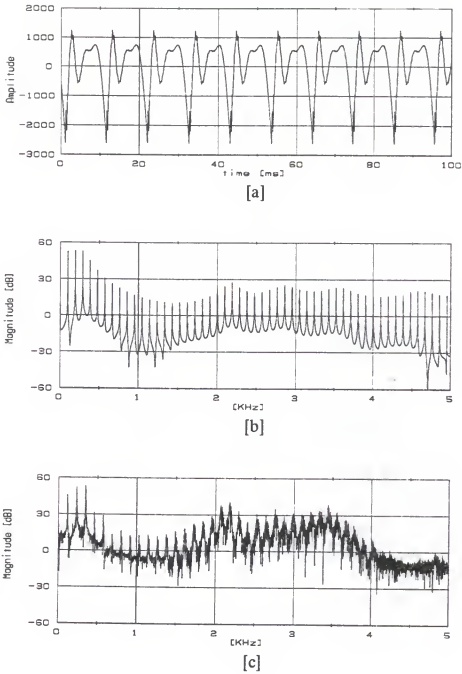
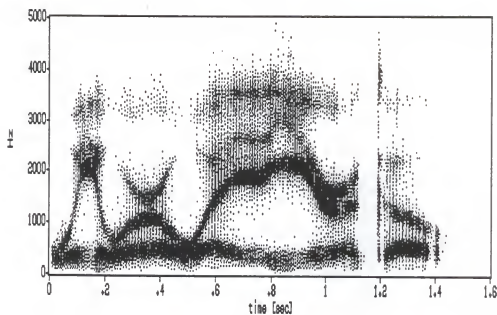
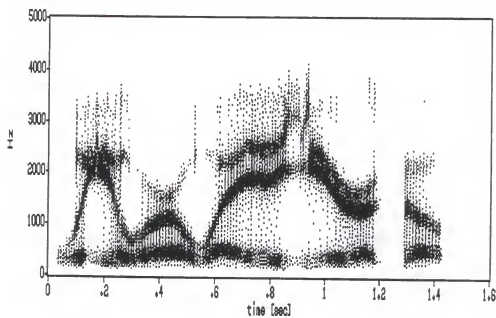


Figure 2-18: Sustained vowel /i/  
 a) Synthesized speech token  
 b) Spectrum of synthesized speech token  
 c) Spectrum of natural speech token



[a]



[b]

Figure 2-19: Spectrogram of sentence "We were away a year ago."  
a) Natural speech  
b) Synthesized speech

### 2.6.2 Example for Specifying Parameters to Flexible Formant Synthesizer

An example for specifying parameters to the flexible formant synthesizer for a sentence “We were away a year ago” is given in Table 2-III and Table 2-IV.

### 2.6.3 Unvoiced Sounds

The excitation source for synthesizing unvoiced sounds is a white-noise source. The white-noise source models the turbulent noise produced by a narrow constriction or an occlusion in the vocal tract during the production of unvoiced sounds. When the constriction in the vocal tract is narrow, the steady air flow from the lungs becomes turbulent, producing friction-like noise which is the source of fricatives (/f/, /θ/, etc.). When the air pressure that builds-up behind an occlusion in the vocal tract is suddenly released, there is a brief interval of friction (due to the sudden turbulence of escaping air) followed by a period of aspiration (steady air flow from the open glottis). The parameter “af” specifies the gain of the noise source during the synthesis of fricatives. The parameter “ah” specifies the gain of the noise source during the synthesis of an aspiration. The synthesizer interpolates the gain of the noise source between the value specified for the previous frame to that specified for the current value over the duration of the current frame. Interpolation of the gain of the noise source provides a more gradual onset and offset of friction and aspiration. However, when a plosive sound has to be generated, i.e., when the value of “ah” or the “af” parameter is suddenly increased by 50 dB from its value for the previous frame, the synthesizer increases the gain of the noise source instantaneously to the specified value for the current frame. Klatt (1980) has mentioned a possibility of adding an exponentially decaying pulse to the noise source at the plosive release time in order to simulate the frequency domain aspects of bursts of air flow due to the sudden release of oral pressure behind the plosive occlusion. In the flexible formant synthesizer, the time

Table 2-III  
Example for specifying parameters for sustained vowels

```
# This specification file is used to synthesize a sustained vowel /t/
#
# Parameter Name      Data type   Value
#
# Sampling rate.
sam_rat              i           10000
# Overall volume control.
g0                   i           112
# Voicing gain.
av                   i           72
# Fundamental frequency.
f0                   i           100
# Pitch synchronous.
PITCH_SYNC           i           1
# Total number of frames to be synthesized.
tot_frames           i           200
# First formant frequency.
f1                   d           302.9
# Second formant frequency.
f2                   d           2172.0
# Third formant frequency.
f3                   d           2851.3
# First formant bandwidth.
b1                   d           133.8
# Second formant bandwidth.
b2                   d           156.5
# Third formant bandwidth.
b3                   d           281.3
# New source model for source excitation and the source spec. file.
src_typ              i           7
# Only the cascade branch will be used.
arch_typ             i           1
# First order FIR filter is used in series with the voicing source.
g_filt              d           -1.0
```

Table 2-IV  
Example for specifying parameters for a sentence

```
# This specification file is used to resynthesize a sentence
# from the analyzed parameter tracks.
# Sentence: "We were away a year ago."
#
# Parameter Name           Data type           Value
#
# Overall volume control.
g0                           d                   92.0
# Fifth formant resonator is not used.
f5                           d000.0              0.0
# Pitch synchronous.
PITCH_SYNC                   i                   1
# Frame size in number of samples when silence.
frame_size                   i                   100
# Only cascade branch is used.
arch_typ                     i                   1
# New source model for source excitation and source specification file.
src_typ                      s7                   src_mod_dmh.d
# Parameter tracks in file dmh_some.d
dmh_some.d                   s                   some
# First order FIR filter is used in series with the voicing source.
g_filt                       d                   -1.0
#
```

Parameter track file: dmh\_some.d

|            |            |             |          |
|------------|------------|-------------|----------|
| f1         | f2         | f3          | f4       |
| d          | d          | d           | d        |
| 0.0        | 0.0        | 0.0         | 0.0      |
| 279.481000 | 390.415000 | 1919.590000 | 3139.630 |
| 277.785000 | 420.984000 | 2127.700000 | 3486.850 |
| 291.245000 | 474.262000 | 2003.430000 | 3300.000 |
| 310.074000 | 582.374000 | 2594.200000 | 3200.000 |
| 301.357000 | 661.551000 | 2646.550000 | 3250.000 |
| 311.055000 | 695.677000 | 2187.430000 | 3214.580 |

constant of the exponentially decaying pulse is specified by the "step\_size" parameter. The duration of the frication and/or the aspiration noise before the onset of the following voiced sounds is defined as the VOT (Voice Onset Time) of the consonant-vowel syllable. The VOT of the desired value can be obtained by specifying the nonzero values of "ah" or the "af" parameter at least for the duration of VOT and then specifying the nonzero value of both the "f0" and "av" parameters.

The typical values of the formant frequencies, bandwidths and amplitudes given in Klatt (1980) for the fricatives and plosives are valid not only for the frication and aspiration portion of these sounds but also serve as "loci" for the formant trajectories of the consonant-vowel transitions. In the cascade/parallel synthesizer configuration, the aspiration portion is generated through the cascade filter bank since the sound source is located at the glottis. The frication portion is generated by using the parallel filter bank. For fricatives, the energy in the frequency range below the first formant is highly attenuated. Therefore, the first formant generator is not excited by the frication noise source and the amplitude control "a1" is set equal to zero dB. For synthesizing the sibilants (/s/, /ʃ/, /z/ and /ʒ/) a sixth formant generator is used to approximate the high-frequency high-energy level in the spectra of these sounds. The spectra of the fricatives (/f/, /v/, /θ/, /ð/) and the plosives (/p/, /b/) do not show any resonance structure. A by-pass path is used in the parallel filter bank to reproduce the flat spectrum of the noise source for these sounds. When synthesizing affricates (plosive release followed by frication noise), the typical values of the formant parameters, given by Klatt (1980), are used to synthesize the frication portion of the affricates. Similarly, when synthesizing the plosives, the typical values of the parameters given by Klatt (1980), are used to synthesize the brief burst of the frication noise generated at the plosive release. These formant frequency values serve as the loci for predicting the formant positions at the onset of the following voiced sounds. During the aspiration portion of the plosives, the formant parameter tracks are

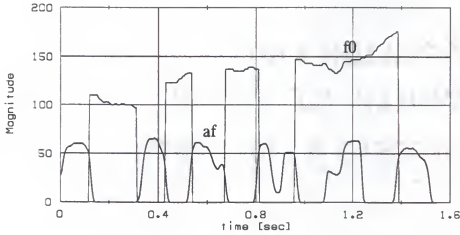


changing to merge with the formant frequencies of the following vowel. The whisper /h/ has the formant parameter values that are the same as those for the following vowel, but the excitation source is aspiration noise. Since the excitation for the whisper is at the glottis, the cascade filter bank should be used for synthesizing a whispery utterance. Figure 2-20a shows the fundamental frequency contour and the frication noise gain contour for the sentence "Should we chase those cowboys?" The formant frequency parameter tracks for this sentence are shown in Figure 2-20b. The spectrogram of synthesized and natural speech utterances of this sentence are shown in Figure 2-21. A visual comparison of the two spectrograms demonstrates a good match of the formant frequency tracks and also of the duration of each sound.

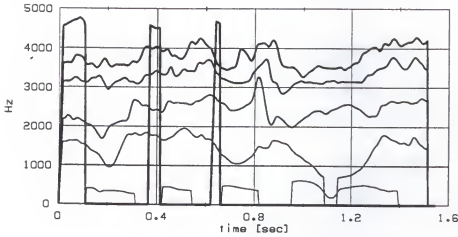
#### 2.6.4 Mixed Excitation Sounds

During the production of mixed sounds (voiced fricatives and plosives) in the human speech production system, the vibrating vocal folds first modulate the airflow from the lungs and a narrow constriction or an occlusion in the vocal tract then produces a turbulent noise from the modulated air flow. Hence, both the glottal source model and the noise source model have to be used in conjunction when synthesizing mixed excitation sounds. In the flexible formant synthesizer, for synthesizing mixed excitation sounds, the parameters, "f0," "av," and "ah" and/or "af", should be nonzero. The noise source is amplitude modulated pitch-synchronously by an amplitude-time waveform as described earlier.

During the interval before the plosive release there is no sound radiated from the lips. However, there is often a small amount of low-frequency energy radiated from the walls of the throat. This low-frequency energy observed in the spectrograms of voiced plosives is called "voice-bar." When synthesizing the "voice-bar," the synthesized speech consists only of the low-energy glottal source pulses.

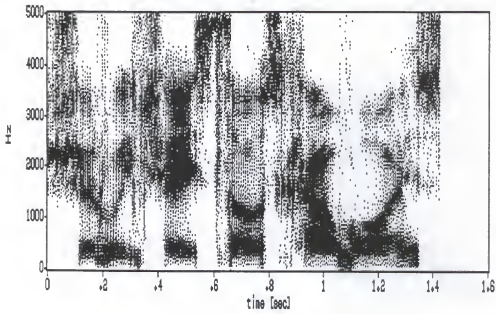


[a]

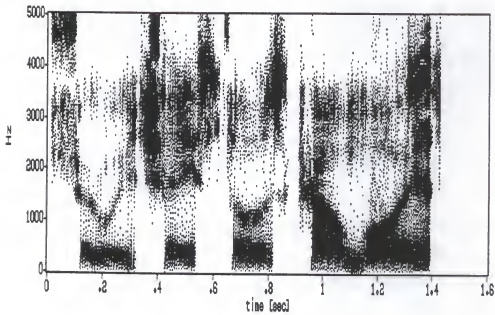


[b]

Figure 2-20: Parameter tracks for the sentence  
 "Should we chase those cowboys?"  
 a) Fundamental frequency and Frication gain contours  
 b) Formant frequency parameter tracks  
 ("f0" is in Hz and "af" is in dB (gain))



[a]

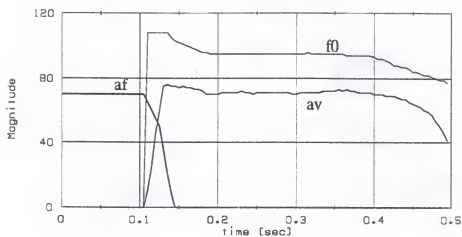


[b]

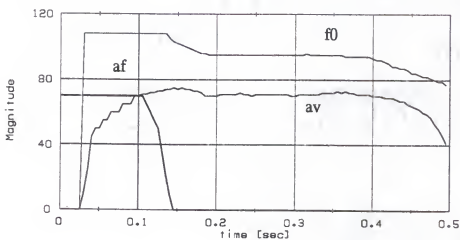
Figure 2-21: Comparison of the spectrogram of the sentence  
"Should we chase those cowboys?"  
a) Natural speech  
b) Synthetic speech

In the case of synthesis of voiced fricatives the glottal source pulses are used to excite the cascade filter bank and the frication noise source is used to excite the parallel filter bank. In the case of voiced plosives both the glottal source and the aspiration noise source are used to excite the cascade filter bank and the frication noise source is used to excite the parallel filter bank. Klatt (1980) has pointed out that the advantage of the cascade/parallel synthesizer configuration is that the synthesis of adjacent voiced sounds can be temporally overlapped with the synthesis of frication sounds, such as for syllables /æʃ/ (as) or /sʊ/ (so), to produce the effect of co-articulation.

When using the all-parallel synthesizer configuration for synthesizing mixed excitation sounds, the first formant generator (resonator) is excited only by glottal source pulses. The second, third, fourth and fifth formant generators (resonators) are excited by a mixture of differentiated glottal pulses and modulated frication (or aspiration) noise. The sixth formant generator and the by-pass path are excited by the frication noise source alone. This strategy, proposed by Klatt [Klatt, 1980] is much simpler than the strategy proposed by Holmes [Holmes, 1983] for producing mixed excitation for formant generators (resonators) in the parallel filter bank. The frication gain contour, the voicing gain contour and the fundamental frequency contour for the speech utterance /feɪl/ are shown in Figure 2-22a. The VOT for this utterance is 110 msec. When the voicing portion overlaps the frication portion at the beginning of the utterance, i.e., the VOT decreases to 30 msec as observed from Figure 2-22b, the same speech utterance is perceived as /veɪl/. The synthetic speech signal and the spectrogram for these two utterances are shown in Figure 2-23 and Figure 2-24

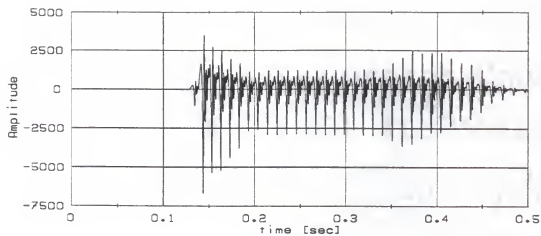


[a]

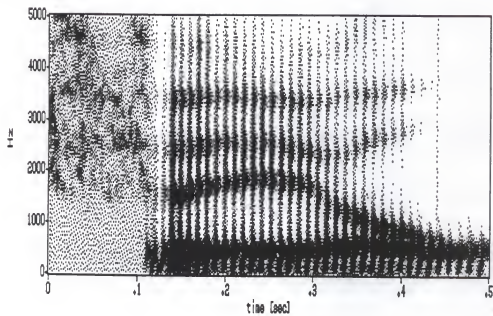


[b]

Figure 2-22: Comparison of “av,” “f0” and “af” parameter tracks for  
 a) Speech token /feil/  
 b) Speech token /veil/  
 (“f0” is in Hz and “af” and “av” are in dB (gain))

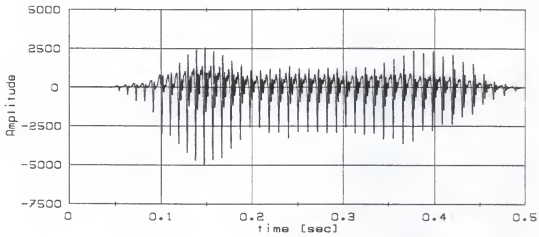


[a]

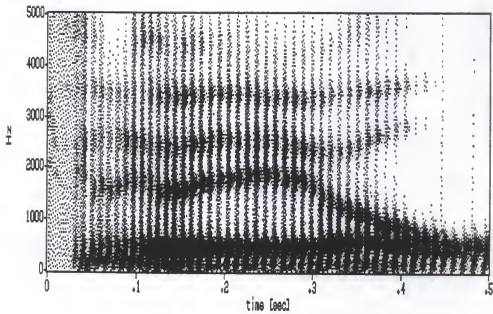


[b]

Figure 2-23: Speech token /feil/  
a) Synthesized Speech token  
b) Spectrogram



[a]



[b]

Figure 2-24: Speech token /verl/  
a) Synthesized Speech token  
b) Spectrogram

## 2.7 Summary

So far we have described the basic features of the flexible formant synthesizer. Many of these basic features are common to Klatt's cascade/parallel formant synthesizer. We have enhanced Klatt's synthesizer by incorporating many new parameters and modifying the synthesis algorithms and the synthesizer architecture. These enhancements have resulted in an increase in the efficiency of the synthesizer. We have improved the specification of: 1) duration of synthesis of a speech utterance, 2) type of synthesis, such as pitch-synchronous or fixed-frame synthesis, 3) first order filters in the synthesizer architecture, 4) variable parameter tracks for synthesis of an utterance, 5) filter specification procedure, etc. We have also developed the FOS for adding flexibility to the synthesizer architecture. Our filter specification procedure allows for configuring the filter banks of the flexible formant synthesizer as required. Also, we have improved Klatt's procedure for simulating the cascade filter bank by a parallel filter bank and also developed an entirely new procedure for the same. We have outlined a few simple strategies for synthesizing voiced, unvoiced and mixed excitation sounds. In the next chapter we discuss some of the advanced features of the flexible formant synthesizer.



## CHAPTER 3 ADVANCED FEATURES OF THE FLEXIBLE FORMANT SYNTHESIZER

### 3.1 Introduction

This chapter describes some of the new features incorporated in the flexible formant synthesizer that are not present in Klatt's cascade/parallel formant synthesizer. First the flexible filter banks are described which allow the user to specify a variable number of filters for the cascade and parallel filter banks at the start-up and also during the synthesis. Then we describe the advantages of a variable sampling rate for synthesis. The procedures to obtain time and/or frequency scaling of the speech signal with or without variable speaking rate are described along with the simulation of source-tract interaction.

### 3.2 Flexible Filter Banks [Lalwani and Childers, 1991a]

Klatt's cascade/parallel formant synthesizer has a rigid configuration (architecture). The user is unable to configure the filter banks according to his or her requirements. The limitations of the filter banks are described in Appendix A. Our design allows the user to configure these filter banks by appropriate filter specifications.

#### 3.2.1 Formant Tracking and Formant synthesis

In the formant synthesizer the formant and anti-formant tracks (time-varying formant and anti-formant frequencies, bandwidths and amplitudes) model the

piecewise time-varying changes in the resonance and anti-resonance characteristics of the vocal tract during an utterance. The formant tracks can be obtained by

- 1) analyzing the natural speech signal [Pinto et al., 1989], and
- 2) concatenating the values of the formant and anti-formant parameters (frequency, bandwidth and amplitude) for each phoneme in the utterance. These values can be obtained from the databases used by rule-based speech synthesis systems [Klatt, 1987].

The number of formant and anti-formant tracks generated by either of these methods will be variable for different utterances. The formant and anti-formant tracks may also be continuous or discontinuous depending upon the sounds being analyzed and/or synthesized. For example, experiments designed to assess the significance of each formant (frequency, bandwidth and amplitude) for the intelligibility and naturalness of various phonemes may involve a variable number of formants tracks. Discontinuous formant tracks may result during the transition from a sibilant sound to a vowel or from a nasal sound to a vowel in an utterance.

In order to be compatible with the speech analysis techniques or the rule-based speech synthesis systems, the formant synthesizer should have

- 1) the flexibility to specify a variable number of continuous and discontinuous formants and anti-formants, and
- 2) a configuration suitable for synthesizing "smooth" speech signals even from discontinuous formant and anti-formant tracks.

A variable number of continuous and discontinuous formants and anti-formants correspond to a variable number of filters in the cascade and parallel filter banks at the start-up and also during the synthesis. The filter bank(s) in both Klatt's cascade/parallel formant synthesizer and Holmes' all-parallel formant synthesizer are incapable of providing such flexibility. Therefore, the user cannot specify a variable

number of formants and anti-formants at the start-up of synthesis and also during the synthesis.

### 3.2.2 Problems with Flexible Configuration of Filter Banks

Several problems arise if the user specifies a variable number of filters in the filter banks at the start-up and/or during the synthesis. Perhaps, these problems may have constrained the development of formant synthesizers with flexible configuration (architecture). Some of the major problems are as follows:

- 1) Changes in the filter bank(s), such as the number of filters or the sequential order of the filters (when arranged in increasing order of the center frequencies) across speech frame boundaries can cause undesired transients in the synthesized speech. These transients may be perceived as “clicks” and “pops” in the synthesized speech.
- 2) An anti-resonator cannot be used to create an anti-formant in the magnitude frequency response of the parallel filter bank. Unlike the cascade filter bank, the anti-formant may not be observed in the magnitude frequency response of the parallel filter bank at the center frequency of an anti-resonator. Also, the skirt response of an anti-resonator has very high amplitude at high-frequencies. Therefore, the magnitude frequency response of an anti-resonator may sometimes greatly affect the overall magnitude frequency response of the parallel filter bank.
- 3) In the all-parallel synthesizer configuration, the parallel filter bank should simulate the magnitude frequency response of the cascade filter bank during the synthesis of voiced sounds [Klatt, 1980]. Klatt (1980) has described a procedure for simulating the magnitude frequency response of the cascade filter bank by the parallel filter bank. As mentioned earlier, this procedure applies only when the cascade and parallel filter banks use exactly five resonators for synthesizing voiced sounds with bandwidth limited to 5 KHz.

### 3.2.3 Solutions to the Problems with the Flexible Configuration of Filter Banks

At the initiation of synthesis (start-up) or during the synthesis of an utterance, whenever there is a change in the number or the sequential order of the filters in the cascade and/or parallel filter banks across the frame boundary, we can create (configure) “new” filter banks with the required number, sequential order and type of filters. This strategy may solve the problem associated with specifying a variable number of filters at the start-up and/or during the synthesis. However, the substitution of the “old” filter bank with a “new” filter bank may cause large transitions in the energy level of the synthesized speech signal. The changes in the energy levels in the synthesized speech signal are perceived by the listeners as undesirable variations in loudness. If the stored energy in the “old” filter bank(s) is allowed to dissipate in the “current” frame while the energy level in the “new” filter bank is increasing, such large transitions in the energy level of synthesized speech signal can be avoided. This method is equivalent to the “overlap-add” method used for filtering a long sequence of data [Oppenheim and Schaffer, 1975], except that the filter coefficients may be changing at each frame boundary. All the cascade (parallel) filter banks, “new” and “old,” can be arranged in parallel to form a “cascade (parallel) branch.” A branch is a parallel configuration of filter banks of the same type (cascade or parallel). When the output of the filter banks in the branch is combined (added), the resulting synthesized speech signal is free from large transitions in the energy level.

We can develop an algorithm that uses the “new” and “old” filter banks as follows:

- 1) The coefficients of the filters in the “old” filter bank(s) are not updated. Its memory is not cleared.
- 2) The coefficients of the filters in the “new” filter bank(s) are calculated based upon the values of the center frequencies and bandwidths of the filters specified for the “current” frame. The “new” filter bank has no memory, i.e., no stored energy.

- 3) The excitation source to the filter bank is switched from the "old" filter bank to the "new" filter bank.
- 4) The "old" filter banks are left "free-running," i.e., the output of an "old" filter bank is obtained from the stored energy in the filter bank and not by driving it with an excitation source.
- 5) The output of the filter banks in the cascade (parallel) branch are added together to obtain the total output of the cascade (parallel) branch.
- 6) The total output from the cascade and parallel branches are added together to obtain the output speech signal, i.e., or the volume-velocity at the lips.

This algorithm generates a smooth synthesized speech signal. However, it requires a large number of computations and a dynamic allocation of a large amount of memory during the synthesis.

After "free-running" for two to three frames, an "old" filter bank contributes a negligible amount of energy to the total output from a branch. Therefore, an "old" filter bank can be excluded from the computation of total output after it has been "free-running" for a certain number of frames, without much affect on the total output of a branch. Consequently, the algorithm can be modified to reduce the number of computations and to reduce the demand for dynamic memory allocation. In this modified algorithm, a fixed number of filter banks,  $N$ , are provided to each of the cascade and the parallel branches. These filter banks do not have any pre-assigned configurations, i.e., no pre-assigned number, sequential order and type of filters. This algorithm has the following characteristics:

- 1) At start-up, a filter bank is configured according to the filter specifications for the first frame. The filter coefficients are calculated and assigned to the filters. Only one filter bank is "active," i.e., producing an output, in each branch.
- 2) At the first frame boundary, the next available filter bank is configured according to filter specifications. The filter coefficients are calculated and assigned to the filters in

the new filter bank. The first filter bank is left “free-running.” The newly configured filter bank is considered as the “current” filter bank and the filter bank that was “current” for the previous frame is considered as the “old” filter bank. The excitation source is connected to the “current” filter bank and is cutoff from the “old” filter bank. The output from the branch is the sum of the output of the two “active” filter banks in parallel.

3) At the  $k^{\text{th}}$  frame boundary, the filter bank which is not “free-running” is reconfigured according to filter specifications for the  $k^{\text{th}}$  frame and its memory is cleared. The filter coefficients are calculated and assigned to the filters in this filter bank. The “old” filter bank in each branch that has been “free-running” for  $(n-1)$  frames ( $n \leq N$ ) becomes “inactive” and stops “free-running” (there is only one such filter bank per branch). This filter bank is available for future use; in other words, it is recycled. The excitation source is connected to the “current” filter bank and is cutoff from the filter bank that was “current” for the previous frame. There are ‘ $n$ ’ “active” filter banks in each branch for each frame (one “current” and  $(n-1)$  “free-running” filter banks) and output of each branch is the sum of the output of the ‘ $n$ ’ filter banks in parallel.

A similar strategy has been used by Verhelst and Nilens (1986) in the “modified-superposition speech synthesizer” to reduce the “clicks” and “pops” in synthesized speech. They have shown that the “clicks” and “pops” are due to large transients in the synthesized speech signal. These large transients arise when the stored energy in the filter bank from the previous frames is dissipated through the filter bank whose coefficients have been abruptly changed across the frame boundary. The abrupt changes in the filter coefficients may be due to large transitions in the formant and anti-formant parameter tracks. They have used a parallel configuration of two cascade filter banks with fixed configurations. Each cascade filter bank is reused with every other frame for synthesizing speech from the updated coefficients and the excitation

source, meanwhile, the other filter bank is left "free-running." They do not check for the possibility of a large transient in the filter bank output before "switching" the filter banks at each frame boundary. We have observed that the transitions in formant and anti-formant frequencies and bandwidths at each frame boundary do not necessarily result in a large transient in the filter bank output. A large transient in the filter output can be detected at the frame boundary from the "initial condition response" of each filter with

- 1) the updated filter coefficients and
- 2) the "old" filter coefficients.

If the ratio of the value of the first sample of the two initial condition responses at the frame boundary exceeds a threshold value, there may be a possibility of a large transient in the output of that filter. A large transient in the output of a filter in the cascade and parallel filter banks may also result in a large transient in the filter bank output.

Calculation of the total output of a branch, even with only 'n' active filter banks, requires considerable computation. If the "current" filter bank is not "switched" frequently, i.e., the "current" filter bank is not changed to the "old" filter bank frequently, the number of "free-running" filter banks will be less than 'n-1' for many frames, and the total amount of computations will be reduced. In fact, when small-amplitude short-duration transients are generated due to small formant transitions, "switching" the filter banks may not be necessary. Also, when synthesizing sustained vowels, switching the filter banks is not necessary at all. Since the configuration of filter banks does not change at each frame boundary and the large transients do not occur at each frame boundary, the number of computations can be further reduced. In the modified algorithm for using multiple filter banks in each branch, the "current" filter bank is "switched" only if

1) There is a change in the number of filters, the sequential order of the filters or the type of filters in the filter bank.

2) There is a possibility of a large transient in the filter bank output.

If, at a frame boundary, any one of the above two conditions is satisfied, the “current” filter bank is “switched.” The “oldest” filter bank with its memory cleared is reconfigured according to the filter specifications for the “current” frame, and the “current” filter bank is left “free-running.” Otherwise, the “current” filter bank is updated, i.e., the coefficients of the filters in the same filter bank are updated according to the filter specifications for the “current” frame. We have modified the flexible formant synthesizer architecture to implement multiple filter banks in each branch. The modified flexible formant synthesizer is shown in Figure 3-1. We have also modified the synthesizer algorithm to incorporate multiple filter banks in each branch.

According to Verhelst and Nilens (1986), the possibility of occurrences of large transients in the parallel filter bank is much less compared to that for the cascade filter bank. However, we have observed large transients in the output of both the filter banks as a result of large transitions in the formant frequencies. Examples of large transients in both the cascade and the parallel filter bank are shown in Figure 3-2. The synthesizer algorithm checks for a possibility of a large transient in the output of both the cascade and the parallel filter banks at each frame boundary. The threshold for detecting a large transient in the output of each filter in the “current” filter bank in the cascade branch is specified by the parameter “*tran\_cas*” and for the parallel branch by the parameter “*tran\_par*.” The user can vary the value of the threshold parameters for the branch that has a large transient in its output, until the “clicks” and “pops” in the synthesized speech are not perceptible. The lower the value of the threshold parameter, the less likelihood of an occurrence of “clicks” and “pops” in the synthesized speech. The ratio of the first samples of the “initial condition responses”



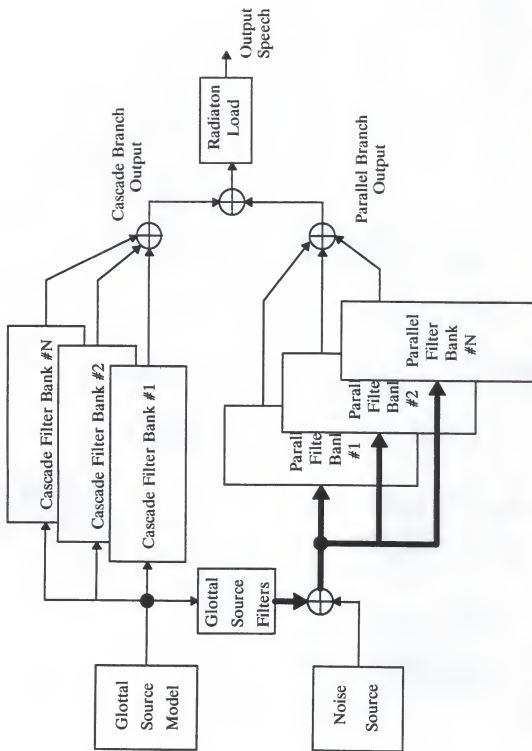


Figure 3-1: Block diagram of the flexible formant synthesizer with multiple filter banks

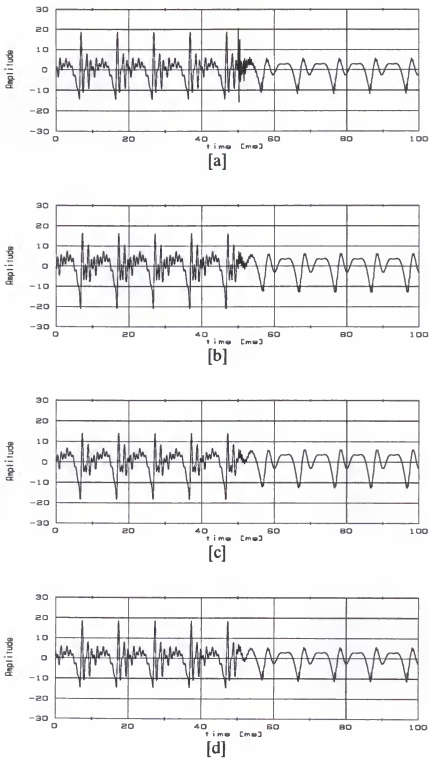


Figure 3-2: Presence of transients ("clicks") in the output from the  
 a) Cascade filter bank  
 b) Parallel filter bank (using Klatt's scale factors and procedure)  
 c) Parallel filter bank (using new scale factors and Klatt's procedure)  
 d) Parallel filter bank (using the new procedure)

of a filter at a frame boundary depends upon the filter coefficients for both the previous frame and the “current” frame and also upon the filter’s initial conditions. The combination of filter coefficients and the filter’s initial conditions preclude the possibility of determining an optimal value for each of the “tran\_cas” and “tran\_par” parameters. For the speech tokens we synthesized, we found that when both thresholds were set equal to 100, the “clicks” in the synthesized speech signal were imperceptible without causing the “current” filter bank to “switch” frequently. In the flexible formant synthesizer, the default values of “tran\_cas” and “tran\_par” parameters are set equal to 100. The user should lower the value of the threshold parameters for the cascade and/or parallel branch, if the “clicks” and “pops” are perceptible. In the traditional approach to remove “clicks” and “pops” from synthesized speech, the user modified (smoothed) the formant tracks until the “clicks” and “pops” were not perceptible. In our approach, the user removes “clicks” and “pops” from synthesized speech by lowering the values of the “tran\_cas” and/or “tran\_par” parameters without having to smooth the formant tracks. Examples are described later.

The second problem with the flexible filter banks is that the anti-resonators cannot be used to create an anti-formant in the magnitude frequency response of the parallel filter bank. The magnitude frequency response of an anti-resonator, normalized to 0 dB at dc, has a skirt response with significant amplitude ( $\gg 0$  dB) at high frequencies. Therefore, the amplitude of other formants in the magnitude frequency response are greatly affected. In the flexible formant synthesizer, we solve this problem by connecting a first order IIR (Infinite Impulse Response) filter in series with each anti-resonator specified for the parallel filter bank in order to attenuate the amplitude of the skirt response of each anti-resonator at high frequencies. The bandwidth of the IIR filter is selected to be larger than the sum of the center frequency and half the bandwidth of the anti-resonator. The parameter “pl\_filt” specifies the filter coefficient of the IIR filter. However, we have achieved only partial success

success with this simple technique and further modifications are required as observed from Figure 3-3. The user can specify the anti-resonator(s) for the parallel filter bank in order to create anti-formant(s) in the magnitude frequency response of the parallel filter bank. However, the anti-resonator(s) are not used for creating anti-formant(s) in the magnitude frequency response of the parallel filter bank. The user can use the automatic procedure (described later) or filter specifications to assign appropriate values to the scale factors for "initial phase" of the output of the resonators in the parallel filter bank (described earlier). The frequency location and the bandwidth of the anti-formants created by this method depend upon the frequency and bandwidth of the formants which created the anti-formants, and hence may not always be as specified.

The third problem deals with the simulation of a cascade filter bank by a parallel filter bank. For the parallel filter bank with only resonators, the formants in the magnitude response of the parallel and the cascade banks should be equal in amplitude during the synthesis of the voiced sounds. Klatt's procedure [Klatt, 1980] and our procedure to achieve this simulation have been discussed earlier. Unlike Klatt's procedure, our procedure does not depend upon the number, sequential order or the type of filters in either the cascade or the parallel filter banks. Also, the synthesizer algorithm employs this procedure at each frame boundary. Therefore, this procedure can be used when the user wants to specify a variable number of filters or vary the sequential order of the filters in the filter banks, at the start-up and also during the synthesis.

Our procedure to automatically match the formant peaks in the cascade and parallel filter banks has been discussed earlier. In order to automatically match the anti-formants, or to create the anti-formants specified for the parallel filter bank only, the flag "PLUS\_MINUS" is set. When the flag "PLUS\_MINUS" is set, the scale factors for the "initial phase" of the output of the resonators in the parallel filter bank are

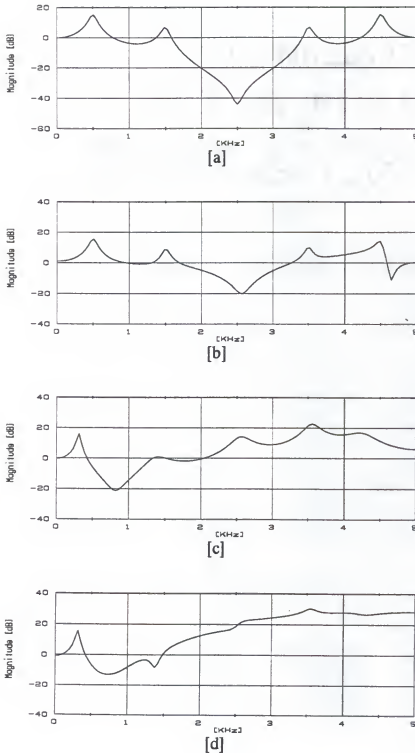


Figure 3-3: An anti-resonator in the cascade and parallel filter bank  
 I) At the third resonance frequency of an uniform tube  
 a) Magnitude frequency response of the cascade filter bank  
 b) Magnitude frequency response of the parallel filter bank  
 II) In between the first and second resonance of an uniform tube  
 c) Magnitude frequency response of the cascade filter bank  
 d) Magnitude frequency response of the parallel filter bank

automatically assigned  $\pm 1$  values. When automatically assigning the value to scale factors for each resonator in the parallel filter bank the following procedure is as follows:

- 1) The resonators in the parallel filter bank are accessed in increasing order of their center frequencies.
- 2) The value of the scale factor for the first resonator (the resonator with the lowest center frequency) is set equal to +1.
- 3) For the rest of the resonators in the parallel filter bank:
  - a) The value of the scale factor for the "current" resonator is set equal to the value of the scale factor of the "previous" resonator, if an anti-formant has to be created in between the formants created by the "previous" and the "current" resonator.
  - b) Otherwise, the value of the scale factor for the "current" resonator is set opposite of that of the "previous" resonator in order to avoid creation of an anti-formant.

The advantage of this method is that it creates formants and anti-formants at appropriate frequencies in the magnitude frequency response even if there are changes in the number of the filters or the sequential order of filters in the cascade and parallel filter banks during the synthesis. If this flag is not set, the default values or the values specified by the user at the start-up are assigned to the scale factors. Therefore, if the flag "PLUS\_MINUS" is not set and the number or the sequential order of the filters changes during the synthesis, the magnitude frequency response of the parallel filter bank may have unwanted anti-formants or may not have the required number of anti-formants.

Our procedure for automatic simulation of the cascade filter bank by a parallel filter bank is accomplished by matching the amplitude of the formants and approximately matching the frequency locations of the anti-formants in the magnitude frequency response of the cascade and parallel filter banks. Our procedure requires more computation than the method proposed by Klatt. However, our method is more

suitable for cascade and parallel filter banks with a variable number of filters at the start-up and also during synthesis.

### 3.2.4 Configuration of the Filter Banks in the Flexible Formant Synthesizer

The filters (resonators, anti-resonators and multipliers) for each filter bank are specified through "filter specifications." At the start-up, the user specifies the information about all the formants, anti-formants and by-pass paths used for synthesizing a particular token through a list of filter specifications. The procedure used to reconfigure the cascade and the parallel filter banks at the start-up and also at each frame boundary is as follows:

- 1) Make a temporary copy of the list of filter specifications for the current frame.
- 2) Update the values of the variable center frequencies in the temporary filter specifications list.
- 3) Remove the filter specifications for the resonators and the anti-resonators with zero center frequencies (when the formant or anti-formant frequency tracks are discontinuous) from the temporary filter specifications list.
- 4) Remove the filter specifications for the resonator and anti-resonator pairs with equal center frequencies and bandwidths from the temporary filter specifications list.
- 5) Rearrange the remaining filter specifications in the temporary filter specifications list in increasing order of their center frequencies.
- 6) For each branch (cascade and parallel) create a list of the filter-numbers (the numbers assigned to the filters during filter specification) of the filters in the "current" filter bank from the temporary filter specifications list.
- 7) For each frame, the "current" filter bank is "switched" if either of the following conditions is satisfied:

- a) The sequential order of the filter-numbers in the list for the "current" frame is not the same as the sequential order of the filter-numbers in the list created for the previous frame (a change in the number of filters in the filter bank is also reflected by a change in the sequential order of filters in the filter bank).
- b) There exists a possibility of a large transient in the "current" filter bank output.
- 8) Configure/reconfigure the "current" filter bank for each branch from the remaining filter specifications in the temporary filter specifications list.
- 9) Calculate the filter coefficients based upon the updated values of the center frequencies and bandwidths and assign them to the filters in the "current" filter bank for each branch.
- 10) Adjust the scale factor for the amplitude control parameter of each filter in the parallel filter bank.
- 11) Adjust the scale factor for the "initial phase" of the output of each filter in the parallel filter bank.

In each branch (cascade and parallel) there is only one filter bank to which the excitation source is connected. The rest of the filter banks in each branch are left "free-running." The output from the "active" filter banks in each branch are combined together to generate the total output of that branch. The total output from both the branches are combined together to generate the total output of the filter banks, which may be the volume-velocity at the lips,  $U(f)$ , or the output synthesized speech.

After implementing this algorithm, we found that switching the "current" filter banks in a branch while the glottal source pulse is nonzero, i.e., in its open phase, causes large transients in the total output of the branch. The reason being that the excitation to the "old" filter bank is abruptly terminated and excitation to the "current" filter bank is abruptly started. This problem may occur in fixed-frame synthesis, where the glottal source pulse may not always have zero value, i.e., be in the "closed-phase," at the frame



at the frame boundary. When pitch-synchronous synthesis is used, it is normal to find the glottal source pulse in its closed-phase at the end of pitch-period, and therefore, this problem is not observed. Verhelst and Nilens (1986) used a pitch-synchronous synthesis procedure with a "modified superposition formant synthesizer," and therefore, did not observe this problem.

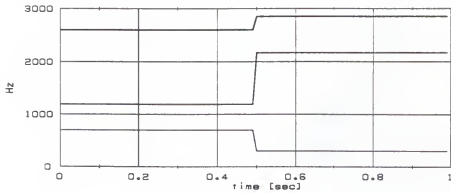
We have modified the above algorithm such that the "switching" of the "current" filter bank in a branch is delayed until the end of the pitch period of the current glottal source pulse. Also, the filter coefficients are not updated until the end of the pitch period of the current glottal source pulse. The synthesis is carried out with the filter coefficients for the previous frame until the end of the pitch period. When the current glottal source pulse is at the end of the pitch period, the "current" filter bank in each branch is "switched" and the filter coefficients are updated. The synthesis is carried out with the updated filter coefficients and the "current" filter bank for the previous frame is left "free-running." If the "switching" of the filter banks at the formant boundary was necessary because of the change in the number and the sequential order of filters in the filter bank, "switching" is necessary even after the delay. However, if the "switching" of the filter banks at the frame boundary was necessary because of the possibility of a large transient in the total branch output, "switching" may not be necessary after the delay (since the initial conditions of the filters are changed). In our algorithm, we do not check if "switching" of the "current" filter bank in each branch is necessary after the delay. (If the delay in "switching" the filter banks is longer than the duration of the "current" frame size, the synthesis will be carried out with the filter coefficients for the previous frame and the filter coefficients for the "current" frame may not be used at all during the synthesis). When the excitation of the "current" filter bank(s) is only the noise source, no such delay is required and the filter bank is "switched" at the beginning of the frame.

We did not observe generation of large transients in the total output of a branch when the “current” filter bank was “switched” even if the noise source was not terminated at the frame boundary. Therefore, the “current” filter bank is “switched” at the frame boundary without any delay when excited by the noise source alone.

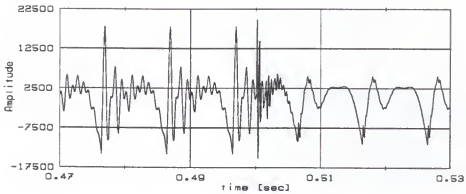
### 3.2.5 Examples

The synthesis algorithm and the architecture of the flexible formant synthesizer enables us to specify a variable number of filters in the filter banks both at the start-up and also during synthesis. A smooth speech signal can be synthesized even from discontinuous formant tracks. The above features of the flexible formant synthesizer are illustrated by the following examples:

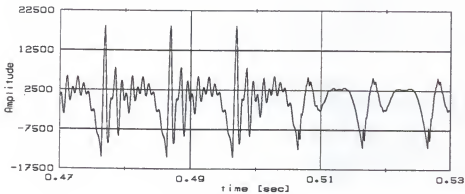
- 1) An abrupt transition in the formant frequencies due to abrupt transition of speech from vowel /a/ to vowel /i/ is shown in Figure 3-4a. Such abrupt transition in the formant frequencies cause a large high-frequency transient in the synthesized speech signal as observed from the Figure 3-4b. A “click” was perceived in the speech signal synthesized using the cascade branch. When the value of the threshold parameter “tran\_cas” was lowered, the “current” filter bank was forced to “switch” at the transition boundary, and the transient in the speech signal disappeared (as observed from the synthesized speech segment in Figure 3-4c) and the “click” was not heard in the synthesized speech signal. The speech signal was synthesized using only the first three resonators (formant generators) in the cascade filter bank and the default values of the fourth and fifth formants were not used.
- 2) An example of using the flexible formant synthesizer to synthesize the sentence “We were away a year ago” from the four unsmoothed formant frequency tracks is shown in Figure 3-5. (Figure 3-5 is similar to Figure 3-4.)
- 3) For the sentence “Should we chase those cowboys?”, the sixth formant track was specified to simulate the high-frequency energy in the unvoiced sounds. Also, for the



[a]

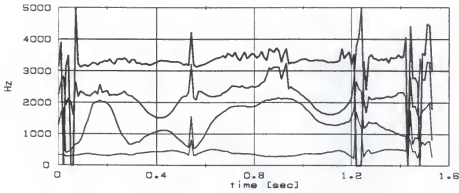


[b]

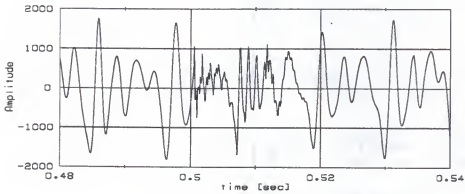


[c]

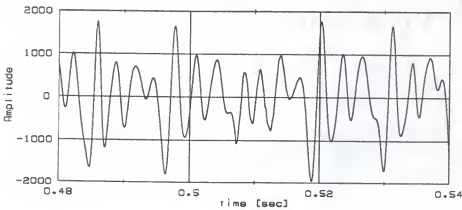
Figure 3-4: Transition from vowel /a/ to /t/  
 a) Formant frequency tracks  
 b) Speech signal in the transition region when synthesized without multiple filter banks  
 c) Speech signal in the transition region when synthesized with multiple filter banks



[a]



[b]



[c]

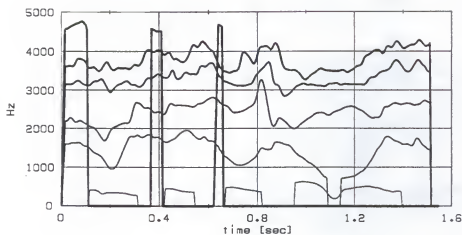
Figure 3-5: Sentence "We were away a year ago."  
 a) Unsmoothed formant frequency tracks  
 b) Speech signal segment corresponding to marked portion of the formant frequency tracks when synthesized without multiple filter banks  
 c) The same speech signal segment when synthesized with multiple filter banks

unvoiced sounds, the first formant was not specified to simulate the attenuation of the energy in the low frequency range for these sounds; however, the first formant was specified for the voiced sounds. The resulting formant tracks, specifically the first and sixth formant frequency tracks, are discontinuous (as observed in Figure 3-6a). With the flexible formant synthesizer, we could synthesize a smooth speech signal even with the discontinuous formant tracks. A portion of synthesized speech signal in a unvoiced/voiced transitions region is shown in Figure 3-6b. Note that the default values of the first and the sixth formants were not used when synthesizing from the discontinuous formant tracks.

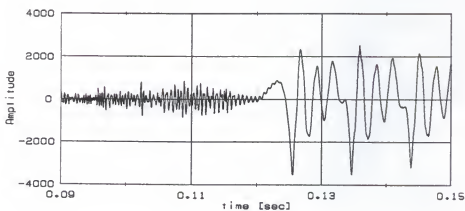
From these figures we can observe that the transients did not occur in the synthesized speech signal when multiple banks were used and the threshold values "tran\_cas" and/or "tran\_par" were lowered to suitable value(s).

### 3.3 Sampling Rate of Synthesized Speech

The bandwidth of a speech signal can be defined as the range of frequency in the spectrum of the speech signal that contains energy above a specified threshold. The bandwidth of synthesized speech is limited to half the sampling rate (Nyquist theorem). Normally, the synthesized speech signal's bandwidth is kept constant by keeping the sampling rate constant during synthesis. The typical bandwidth of 5 KHz for a sampling rate of 10 KHz produces intelligible speech for most voiced sounds but is not adequate for synthesizing plosives and fricatives [Klatt, 1980; Holmes, 1983]. For example, the spectrogram of natural utterances of sibilants (/s/ and /ʒ/) show most of the energy concentration in the frequency range above 5 KHz. Holmes et al., (1990) have recently increased the bandwidth of the all-parallel formant synthesizer to 8 KHz (sampling rate of 16 KHz) in order to improve the quality of the synthesized speech. Specifying a very high sampling rate to synthesize these sounds in an utterance may



[a]



[b]

Figure 3-6: Discontinuous formant frequency tracks for the sentence "Should we chase those cowboys?"  
 a) Formant frequency tracks  
 b) Synthesized speech segment corresponding to marked portion of the formant frequency tracks when synthesized with multiple filter banks

not always be economical and not required for intelligibility of other sounds in that utterance. If the sampling rate could be changed during the synthesis, it might be possible to synthesize these sounds at a higher sampling rate and other sounds at a lower sampling rate. If the sampling rate is varied during the synthesis, we may be able to synthesize all the sounds in an utterance with high intelligibility and yet keep a low sampling rate during the synthesis of most of the utterance. However, the currently available D/A converters do not permit playback of sampled speech signals with a variable sampling rate. Another drawback of a variable sampling rate is that sudden shifts in the signal bandwidth may cause undesirable changes in the energy levels of the higher order formant generators (resonators). Such changes may lead to "clicks" and "pops" in the synthesized speech [Klatt, 1980].

In the flexible formant synthesizer, we have made provision for specifying a variable sampling rate. However, the sampling rate should be kept constant when the synthesized utterance has to be played back through a D/A converter. If the sampling rate of the synthesized utterance has to be decreased or increased, it should be changed for the complete utterance by using decimation or interpolation techniques, respectively [Rabiner and Schafer, 1978].

### 3.4 Time and Frequency Domain Scaling

The time and frequency domain scaling of an utterance involves changing the duration of an utterance, changing the signal bandwidth, and modifying the formant parameters (frequencies and bandwidths) with or without changing the speaking rate. The applications of time and frequency domain scaling are:

- 1) varying the duration of the recorded speech signal,
- 2) varying the speaking/information rate of the speech signal for language teaching and for speed reading for the blind,

- 3) improving the quality of synthesized speech by experimenting with the duration of the phonemes,
- 4) for adjusting the speaking rates (duration of test tokens) in IWR (Isolated Word Recognition) Systems,
- 5) for telecommunication over a narrow baseband channel,
- 6) voice conversion e.g., male to female and vice versa, or for creating new voice, and
- 7) creating novel sounding utterances for toys.

The modification of duration of a speech utterance using a phase vocoder was demonstrated as early as 1966 [Flanagan and Golden, 1966]. Several frequency compression/expansion techniques have been developed for efficient communication of speech signals. Most of the techniques for changing the bandwidth of the speech signal also involve modification of the speaking rate, which is desirable in some applications but not in all. Malah (1979) proposed a system for harmonic bandwidth reduction and time scaling of the speech signal using pitch (fundamental frequency) information. He proposed the application of his system to reduce the signal bandwidth in vocoders and for time-alignment of speech tokens in the IWR systems. Seneff (1982) proposed a speech analysis-synthesis system to independently modify the excitation and/or the magnitude frequency response of the filter banks without explicit pitch extraction. Recently, Moulines and Charpenter (1990) have described a text-to-speech synthesis system, where the time scale modification could be performed either in combination with pitch scaling or independent of pitch scaling. In this system, the pitch contour was scaled to modify both the pitch and the time scale. The time scale could be changed independent of the pitch scale by repeating speech signal from the previous pitch-period and/or by eliminating (cutting) the signal for the entire pitch-period duration. d'Alessandro (1990) has described an analysis-synthesis system, in which, the speech signal was represented by elementary waveforms. He claimed that time and frequency domain modifications could be easily



performed by varying appropriate parameters of this system that were chosen from an acoustic point of view. He classified time and frequency domain modifications of the speech signal into two categories: localized modifications and global modifications. Example of localized modifications are formant parameter modifications, friction noise modifications (duration and spectrum), plosive burst noise modifications (duration and spectrum), etc. Examples of global modifications are pitch modifications, frequency (signal bandwidth) compression/expansion, etc. Flanagan's and Seneff's systems have a "direct approach," in which, the speech signal may be modified by changing the sampling rate (decimation or interpolation) and the spectrum of the speech signal. Other systems have a "vocoder approach," in which, the parameters of the speech synthesis model are extracted from the speech signal using the analysis techniques and the modified speech signal is generated from the synthesis model using the modified parameters.

While developing the flexible formant synthesizer, we designed it so that the user could modify the time and frequency domain characteristics of the signal. Using the "direct approach" to modify the synthesized speech signal seemed redundant. There are several advantages to modifying the synthesizer parameters before synthesis. The formant synthesizer parameters, such as fundamental frequency, formant frequencies, etc., are highly correlated to the acoustical aspects of perception. Any modifications to these parameters result in modification of the corresponding time and frequency domain characteristics of the synthesized speech signal. The effects of these modifications can be perceptually evaluated or visually checked from the spectrograms. The parameters of the synthesizers can be repeatedly modified until the desired time and frequency domain characteristics are observed in the synthesized speech signal. The manner by which the time and frequency domain characteristics of the synthesized speech can be varied are described below.

### 3.4.1 Frequency Compression/Expansion

The frequency domain modifications are normally brought about by modifying the formant parameters (formant frequencies, bandwidths and amplitudes). The advantage of using a formant synthesizer is that these parameters can be directly specified to the synthesizer. The signal bandwidth compression/expansion can be brought about by appropriate scaling of all the formant frequencies and formant bandwidths. (Expansion of signal bandwidth may require an increase in the sampling rate to avoid aliasing problems.) An example of compression/expansion of signal bandwidth is compression of the spectrum of the speech signal at the transmission end of a communication channel and expansion of the spectrum of the speech signal at the receiver end of the channel. The spectrum of the speech signal is compressed by formant frequency scaling to map the high-frequency information to the low frequency range. The sampling frequency can, therefore, be decreased to lower the bit rate. The spectrum of the transmitted signal is expanded and the sampling frequency is increased at the receiver end to generate the original speech signal at the original "speaking rate." The frequency domain modifications may also involve changing some of the formant parameters within their formant region (within their "normal formant parameter range") without changing the signal bandwidth. An example of frequency compression/expansion within the formant region is conversion of a male voice to a female voice, or vice versa, by scaling some of the formant parameters.

### 3.4.2 Time Scaling/Variable Speaking Rate

The time scaling for variable speaking rate can be brought about by

- 1) playback of the speech samples at a sampling rate different from the rate of the synthesis of the speech samples,
- 2) changing the frame size prior to resynthesis of the speech utterance, and

3) altering the parameter update sequence (sequence of values of the variable parameters in their parameter tracks) during resynthesis.

#### 3.4.2.1 Changing the sampling rate

The simplest method to change the time scale and thus the speaking rate of an utterance is to playback its samples at a sampling rate that is different from the original sampling rate. However, changes in the sampling rate are accompanied with the changes in the fundamental frequency, signal bandwidth, formant frequencies and formant bandwidths. These changes may alter the perception of the phonemes in the utterance. The sampling rate for speech can be changed by using decimation and interpolation techniques without altering the perception of the phonemes in the utterance. However, these techniques do not change the time scale and the speaking rate of the speech signal. The time scale or speaking rate of synthesized speech signal can be modified using a phase vocoder [Flanagan and Golden, 1966; Seneff, 1979]. This method is unrelated to the development and applications of the flexible formant synthesizer, and hence is not discussed here.

#### 3.4.2.2 Changing the frame size

A speech frame is the duration of a portion of speech signal for which the time and frequency domain characteristics of the speech signal are assumed to remain constant. Generally, the speech signal is analyzed on a frame-by-frame basis to extract the parameters required for resynthesis. For fixed-frame analysis, the analysis frame size remains fixed during the analysis of the entire speech signal. For variable-frame analysis, the frame size may change during analysis based upon the variations in some criteria, such as fundamental frequency (pitch-synchronous analysis) or short-time

energy in the signal [Pinto et al., 1989]. Changes in the time scale and the speaking rate of an utterance may be brought about by

- 1) scaling the value of the frame size parameter specified for the fixed-frame analysis prior to fixed-frame resynthesis of an utterance,
- 2) scaling the fundamental frequency contour prior to pitch-synchronous resynthesis of an utterance. This method also changes the pitch of an utterance, and
- 3) performing the fixed-frame analysis and the variable-frame resynthesis (by setting the “PITCH\_SYNC” flag for pitch-synchronous synthesis or by specifying a parameter track with varying values for the “frame\_size” parameter), or performing the variable-frame analysis and the fixed-frame resynthesis of an utterance (by specifying the “frame\_size” parameter to be constant).

In the flexible formant synthesizer, modifications to the frame size can be easily specified. An example of time scaling and variable speaking rate for the sentence “We were away year ago” by scaling the fundamental frequency contour prior to the pitch-synchronous synthesis is shown in Figure 3-7.

#### 3.4.2.3 Changing the parameter update sequence

The early experiments with variable speaking rate involved cutting and/or splicing of the speech sounds in order to decrease or increase the time duration and speaking rate of each sound in the utterance. This procedure can be simulated during the synthesis of an utterance using the flexible formant synthesizer. The cutting and splicing of a portion of a sound is equivalent to skipping (decimating) and repeating (interpolating) samples of the variable parameter tracks of an utterance, respectively. The *n*th sample (value) in a parameter track corresponds to the value of that variable parameter for the *n*th frame. If a portion of each of the parameter tracks extracted from the utterance is not used during resynthesis, the corresponding portion (frames) of an utterance is not synthesized. If each value from a portion of each variable

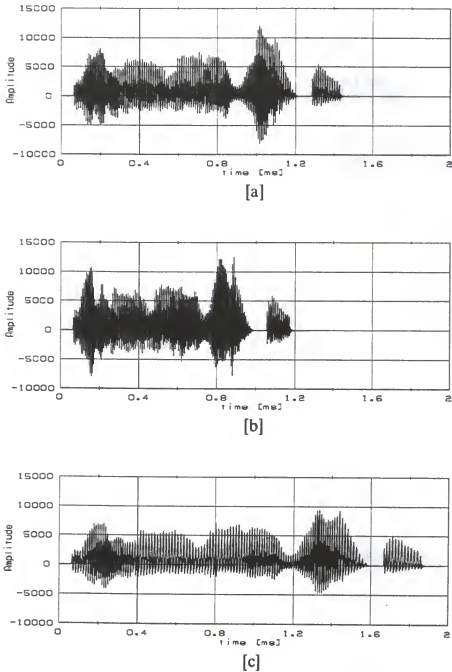


Figure 3-7: Time scaling and variable speaking rate for sentence "We were away year ago" by scaling the fundamental frequency contour  
 a) Normal speaking rate  
 b) Fast speaking rate when  $f_0$  is uniformly scaled by 1.25  
 c) Slow speaking rate when  $f_0$  is uniformly scaled by 0.75

parameter track is repeated during the resynthesis, the duration of the corresponding portion (frames) of an utterance in the synthesized speech is extended. The advantage of using the synthesizer is that a smooth waveform is synthesized by the synthesizer and there are no abrupt transitions in the speech signal due to cutting and splicing of its portions.

During normal synthesis, the variable parameters are assigned values following the serial order of the samples (values) of the variable parameters in their parameter tracks. Changes in the time scale and/or speaking rate can be brought about by altering the serial order of the assignment of values from the parameter tracks to variable parameters to skip frames and/or repeat each frame during the synthesis of portion(s) of an utterance. In the flexible formant synthesizer, either the parameter "tot\_frames" or the parameter "start\_frame" can be used to alter the sequence of the assignment of values from the parameter tracks, i.e., sequence for updating the values of the variable parameters during the synthesis. When synthesizing sustained phonations, i.e., when not a single variable parameter track is specified, the parameter "tot\_frames" determine the total number of frames to be synthesized. If any variable parameter tracks are specified for synthesizing an utterance, the total number of frames that can be synthesized is determined by the length (number of samples) of the parameter tracks. In such a case, the parameter "tot\_frames" (actually, its parameter track) can be used for specifying which values of the variable parameters from the parameter tracks should be repeated or skipped during synthesis. If the value of the parameter "tot\_frames" for the current frame is -1, the previous frame is synthesized (repeated) again. The values of all the parameters (except the parameter "tot\_frames") for the previous frame are repeated for the current frame and the values of all the parameters (except the parameter "tot\_frames") for the current frame may be used for the next frame. If the value of the "tot\_frames" parameter for the current frame is -2, the current frame is not synthesized (skipped). The next frame is considered as

the current frame and the values of all the parameters (except the parameter “tot\_frames”) for the current frame may not be used at all during the synthesis. If the value of the “tot\_frames” parameter for the current frame is 0, the current frame is synthesized with no skipping or repetition. By carefully creating the parameter track for the “tot\_frames” parameter, i.e., by creating an appropriate array of 0, -1 and -2, repetition and/or skipping of the portions of an utterance can be achieved during synthesis. An example of time scaling and variable speaking rate for the sentence “We were away a year ago” by repeating/skipping the speech frames is shown in Figure 3-8. The spectrograms of the speech signals in Figure 3-8 are shown in Figure 3-9.

The parameter “start\_frame” can also be used for altering the parameter update sequence. The parameter “start\_frame” is normally used to indicate the starting frame number, i.e., the starting value of the variable parameters in their parameter tracks that begin the synthesis of an utterance. The rest of the utterance is generated by synthesizing each frame, after the starting frame, in serial order of the sequence of the values of the variable parameters from their parameter tracks. The parameter “start\_frame” can also be used to alter the serial order of a sequence of the values of the variable parameters in their parameter tracks during the synthesis. If the “start\_frame” parameter is specified as a variable parameter, its parameter track determines the serial order in which the frames are synthesized, i.e., the serial order for updating the value of the variable parameters. This method can be used, to change not only the time scale and/or speaking rate, but also to create scrambled utterances for applications, such as in voice security, toys, etc.

Either the “start\_frame” or the “tot\_frame” parameters can be used to modify time scale/speaking rate of an utterance (using both may lead to undesirable changes, unless carefully used). The “start\_frame” or the “tot\_frame” parameter, whichever is specified, is assigned values from its parameter track in a serial order. Other variable parameters are assigned values from their respective parameter tracks in the order

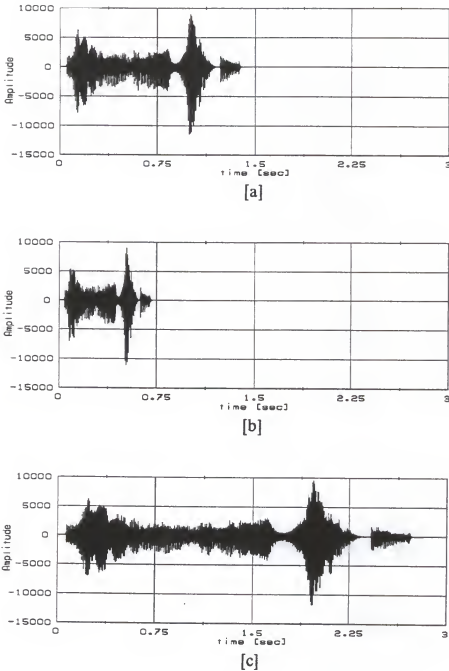


Figure 3-8: Time scaling and variable speaking rate for sentence "We were away year ago" by repeating/skipping the speech frames. The speech signal for

- Normal speaking rate
- Fast speaking rate obtained by skipping alternate frames
- Slow speaking rate obtained by repeating each frame



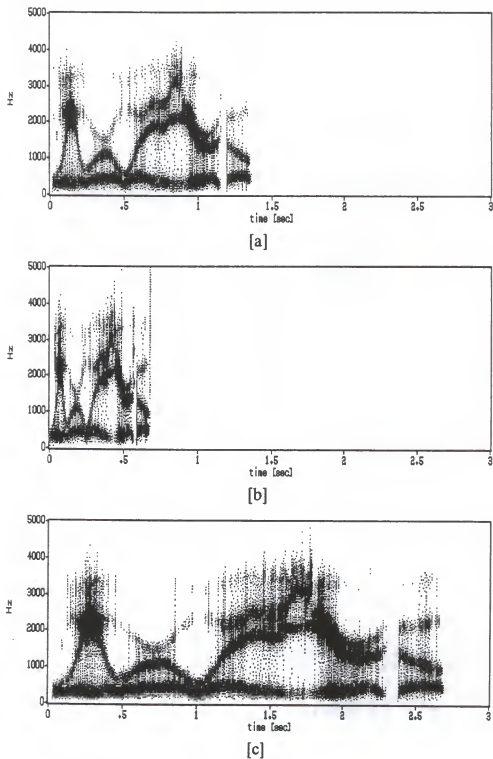


Figure 3-9: Time scaling and variable speaking rate for sentence "We were away year ago" by repeating/skipping the speech frames. The spectrograms of speech signal for  
 a) Normal speaking rate  
 b) Fast speaking rate obtained by skipping alternate frames  
 c) Slow speaking rate obtained by repeating each frame

specified by the “start\_frame” or the “tot\_frame” parameter track. Therefore, the total number of frames to be synthesized is determined by the length of the “start\_frame” or the “tot\_frame” parameter track and not by the length of other parameter tracks.

When using the flexible formant synthesizer, the time-scale and/or speaking rate can be modified by specifying only one additional parameter track (the “start\_frame” or “tot\_frame” parameter track). There is no need to create multiple copies of all the variable parameter tracks with only slight variations.

### 3.5 Source–Tract Interaction

In the conventional speech production model, the source and the vocal tract system are considered to be independent of each other (i.e., varying the vocal tract configuration will not have any effect on the source), which is true when the glottis is closed, or almost closed. But when the glottis is open, interaction between the vocal tract system and the source occurs. This is often called “loading.” It has been shown that the loading of the vocal tract can have an appreciable effect on the glottal flow pulse shape. This is called source–tract interaction. Source–tract interaction has been conjectured to be important for synthesizing high-quality, natural sounding speech. Speech synthesized with source–tract interaction sounds more natural than speech generated without such interaction [Childers et al., 1983; Pinto et. al., 1989]. The effects of source–tract interaction may be observed in the glottal waveforms obtained by inverse filtering of speech [Wong and Markel, 1977; Wong, 1991]. Of the several source–tract interaction phenomena observed from the inverse filtered speech [Wong, 1991], the two most important source–tract interaction phenomena are glottal pulse skewing and the truncation of the first formant.

### 3.5.1 Glottal Pulse Skewing

The glottal flow waveforms obtained by inverse filtering the speech signal often show unsymmetrical glottal pulses, with a slowly-rising glottal opening phase and a sharply terminating glottal closing phase (Holmes, 1962; Rothenberg, 1973; Lee, 1988). Such unsymmetrical glottal pulses may be an important determinant of voice quality affecting the high frequency energy of the waveform, and hence, affecting the levels of the formants. The degree of skewness differs for different vowels, and it has been shown that the first formant load is the most important in determining the degree of skewness [Fant and Ananthapadmanabha, 1982].

### 3.5.2 Truncation of First Formant

It is generally realized that there can be appreciable first formant energy absorbed by the glottis during the open phase of the glottal cycle, when the glottal impedance is finite. This is called truncation, which means the termination of formant oscillations by excessive damping within the glottal open phase [Fant and Ananthapadmanabha, 1982]. Glottal damping causes a truncation of formant amplitudes, changes the formant frequencies and increases formant bandwidths during the glottal open interval. Its equivalent effect is to cause first formant oscillations on the glottal pulses which are observed as ripples in the open phase of the glottal pulses. The main perceptual effect of truncation is a reduction of the loudness level of the formant. The first formant load has been found to be the most important factor for determining the degree of this effect [Fant and Ananthapadmanabha, 1982].

### 3.5.3 Tracheal Poles and Zeros

The resonance characteristics of the trachea and the bronchi are not modelled in the source-filter model of speech production. However, these resonance

characteristics may produce additional formant and anti-formant pairs in the speech spectra of the sounds for which the vocal folds are presumably open during phonation. Fant et al. (1972) and Cranen and Boves, (1987) have measured the values of the lowest three tracheal resonances to be 510, 1350 and 2290 Hz for the male voice and slightly higher for the female voice. Recently, Ananthapadmanabha and Fant (1982) and Rothenberg (1985) have reported that the actual effect of the tracheal resonances on the vocal-tract transfer function is a complex function of the glottal configuration over time. The effect of tracheal coupling can be modelled by adding the most significant "pole-zero" pairs to the vocal-tract transfer function (i.e., by adding one or more resonator and anti-resonator pairs to the filter banks) [Fant et al., 1972; Klatt, 1986].

#### 3.5.4 Simulation of Source-Tract Interaction with the Formant Synthesizer

When simulating source-tract interaction in a source-filter model [Fant, 1960] one must decide which effects of source-tract interaction should be attributed to the glottal source model and which effects should be attributed to the filter bank(s). Normally, the effect of loading of the vocal tract on the glottal source pulse is attributed to the glottal source pulse shape. The effect of truncation of the first formant oscillations due to glottal impedance is attributed to the first formant amplitude.

Ananthapadmanabha and Fant, (1982) have developed a glottal source model that simulates both the time-varying glottal impedance and vocal-tract load to generate the glottal source pulses with right-skewness and a ripple component. Provision has been made to include this model in the flexible formant synthesizer.

The LF model can produce glottal source pulses with right-skewness during the glottal open phase. The degree of skewness can be controlled by the LF model's time-domain parameters. Several other parametric glottal source models incorporated in the flexible formant synthesizer generate left-skewed pulses. The magnitude frequency response of the right or left skewed pulses may be similar but

their phase responses are different. A typical glottal flow pulse obtained from the inverse filtering of the speech signal and the approximation of this glottal flow pulse by a glottal source pulse generated by the LF model is shown in Figure 3-10.

The truncation of the first formant oscillations, i.e., reduction in the first formant amplitude, is simulated by modifying the first formant frequency and bandwidth during the open phase of the glottal source pulses. In the flexible formant synthesizer pitch-synchronous synthesis procedure should be used when simulating source-tract interaction. The values of first formant frequency and bandwidth specified for the current frame are used to calculate the filter coefficients of the first formant generator (resonator) for the closed phase portion of the current glottal source pulse. Prior to the open phase of the following glottal source pulse, the value of the first formant frequency and/or the first formant bandwidth parameters are multiplied by a scale factor (normally to increase the value). These scaled values of the first formant frequency and bandwidth parameters specified for the current frame are used to calculate the filter coefficients of the first formant generator (resonator) for the open phase of the following glottal source pulse. The values of the first formant frequency and bandwidth specified for the next frame are used to calculate the filter coefficients of the first formant generator (resonator) for the closed phase portion of the following glottal source pulse.

The parameter "op" specifies the open phase duration of a glottal source pulse as a fraction of the pitch-period. The parameter "st\_freq" specifies the value of the scale factor for the first formant frequency and the parameter "st\_bw" specifies the value of the scale factor for the first formant bandwidth. The parameters "ST\_FRAMES" and "ST\_STEP" are flags to indicate how the values of the first formant frequency and bandwidth parameters should be changed to their scaled values during the open phase of the glottal source pulse. If the flag "ST\_FRAMES" is set, the values of the first formant frequency and bandwidth parameters are changed

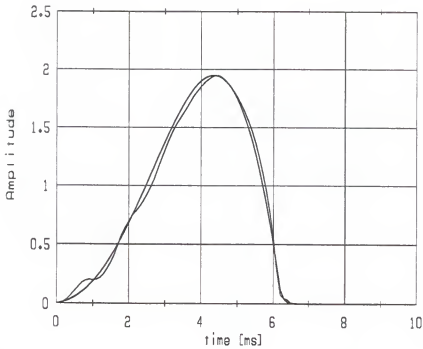


Figure 3-10: An illustration of right-skewness of the glottal flow pulse obtained by inverse filtering of the speech waveform, and its simulation with a glottal source pulse generated by the LF model.

abruptly at the beginning of the open phase, remain constant for the entire duration of the open phase and then change abruptly to the values specified for the following glottal source pulse. If the parameter "ST\_STEP" is set, the values of the first formant frequency and bandwidth parameters are incrementally changed to the scaled values for the first half of the open phase and then incrementally changed to the values specified for the following glottal source pulse for the last half of the open phase. These two methods for changing the first formant bandwidth and frequency during the open phase of the glottal source pulses are described in Figure 3-11.

The truncation of first formant oscillations due to increase in the first formant bandwidth in the open phase portion of the glottal source pulses can be observed in Figure 3-12 and Figure 3-13. It can be observed that both methods of changing the values of the first formant bandwidth causes truncation of first formant oscillations in the open phase portion of the glottal source pulses. However, changing the first formant bandwidth abruptly to the scaled value causes an undesirable high-frequency ripple in the output signal. Such undesirable ripple is not observed when the first formant bandwidth is incrementally changed to the scaled value during the open phase portion of the glottal source pulses. To achieve the same effect, the value of the scaling factor required when the first formant bandwidth is incrementally changed is double the value required when the first formant bandwidth is abruptly changed. The effect of changing (increasing) the first formant frequency during the open phase portion of the glottal source pulses is shown in Figure 3-14. The higher the value of the scale factor, the higher is the frequency of the first formant oscillations in the portions of the output waveform corresponding to the open phase of the glottal source pulses. However, no truncation of the first formant oscillations is observed. A comparison of the output waveforms in Figure 3-12 through Figure 3-14 show that changing the first formant bandwidth incrementally during the open phase portion of the glottal source pulse may be the best way to simulate truncation of the first formant oscillations.

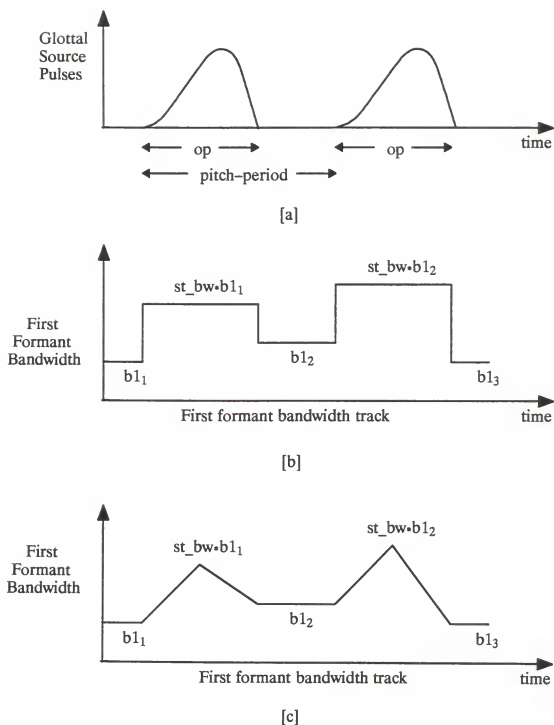


Figure 3-11: Two methods for varying the first formant bandwidth and/or frequency during the open phase of the glottal source pulses.  
 a) Glottal source pulses  
 b) When the flag "ST\_FRAME" is set  
 c) When the flag "ST\_SMP" is set



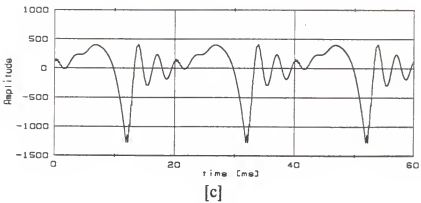
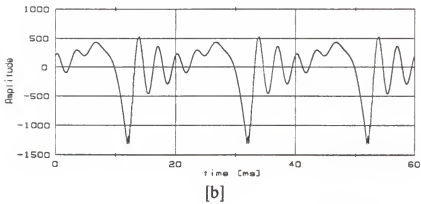
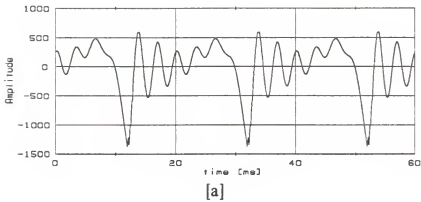


Figure 3-12: Truncation of first formant by increasing the first formant bandwidth

- a) No change
- b) Increased incrementally to twice its specified value
- c) Increased abruptly to twice its specified value

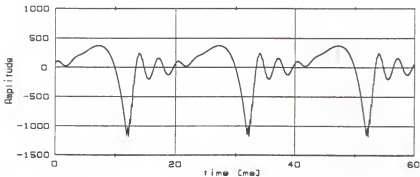
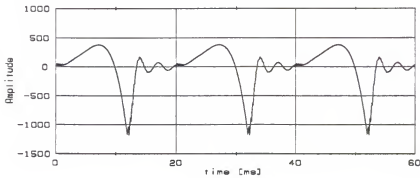
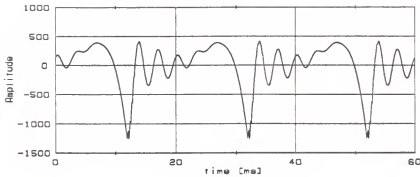


Figure 3-13: Truncation of first formant by increasing the first formant bandwidth during the open phase

- a) Increased incrementally to four times its value
- b) Increased abruptly to four times its value
- c) Increased incrementally to eight times its value

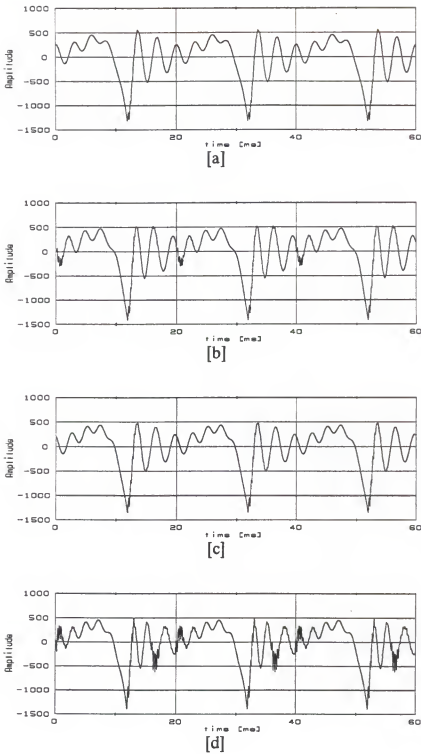


Figure 3-14: Simulation of source-tract interaction by increasing the first formant frequency

- a) Increased incrementally to 1.2 times its value
- b) Increased abruptly to 1.2 times its value
- c) Increased incrementally to 1.5 times its value
- d) Increased abruptly to 1.5 times its value

In the recent version of Klatt's cascade/parallel synthesizer [Klatt and Klatt, 1990], similar features have been incorporated to simulate source-tract interaction. He has incorporated the LF model and modified his model to provide optional glottal source models. Both models can generate right-skewed glottal source pulses. Also, the values of the first formant frequency and bandwidth parameters can be abruptly changed during the open phase by a scale factor. However, Klatt's synthesizer does not have an option for changing the values of the first formant frequency and bandwidth parameters incrementally.

Also, Klatt's cascade/parallel synthesizer has a provision for adding a tracheal resonator and anti-resonator pair in series with the cascade filter bank and for adding a single tracheal resonator to the parallel filter bank (similar to the representation of the nasal tract). This resonator and anti-resonator pair simulates the most significant tracheal formant and anti-formant pair that is significant for synthesis [Klatt and Klatt, 1990]. A two step procedure for including the tracheal "poles" and "zeros" in the vocal-tract transfer function for synthesis of some sounds: 1) the values of the center frequencies of the tracheal resonators and anti-resonator should be moved together to the desired "pole" frequency, and 2) the value of the center frequency of the tracheal anti-resonator should be gradually moved to another value obtained from spectral analysis or parameter databases. They have suggested another strategy, in which, the center frequencies of the tracheal resonators and anti-resonators are normally kept overlapped, and the bandwidth of the tracheal anti-resonator is increased (and/or the bandwidth of tracheal resonator is decreased) in order to reveal the presence of a tracheal formant when required.

In the flexible formant synthesizer, we have made provision for specifying additional filters to the default configuration of the cascade and parallel filter banks. The user can specify these additional filters as tracheal resonators and anti-resonators in the filter banks by using the additional filter specifications available to the user.

### 3.6 Summary

In our software, the procedure for specifying filter parameters, the flexible synthesis algorithm and the flexible synthesizer architecture enables us to create a flexible configuration of the cascade and the parallel filter banks. The dynamic configuration of the filter banks at the start-up and also at each frame boundary enables us to have a variable number of filters in the cascade and parallel filter banks at start-up and during synthesis. With the flexible formant synthesizer, we can synthesize speech from the exact number of continuous and discontinuous formants and anti-formants specified for synthesis (without inserting default values for the formants and anti-formants that are not specified). The synthesized speech is free of “clicks” and “pops” even if the formant tracks are unsmoothed and/or discontinuous.

We have incorporated algorithms for time and frequency scaling of the synthesized speech signal. In the flexible formant synthesizer, we have provided parameters appropriate for time and frequency scaling of the speech signal. The flexible formant synthesizer synthesizes a smooth speech signal even if the formant tracks are abruptly changed during the skipping and/or repetition of the frames.

In the flexible formant synthesizer, we have made provision for simulating source-tract interaction. A glottal source model, such as the LF model, can be used to generate glottal source pulses with a specified skewness. The first formant bandwidth and frequency can be changed during the open-phase of the glottal source pulse to simulate the effect of truncation of the first formant oscillations during the open-phase portion of the glottal source pulses. We have observed that changing the first formant bandwidth incrementally provides a better simulation of the first formant truncation effect than other methods.

## CHAPTER 4 GLOTTAL SOURCE MODEL

### 4.1 Introduction

In this chapter we describe a new unified glottal source model developed for modeling and synthesizing various vocal characteristics and vocal disorders, such as modal, creaky, breathy, rough and hoarse voice. First, we give a brief review of the glottal flow characteristics that are known to be significant for modeling these vocal characteristics. Then we describe a new glottal source model, which can simulate the glottal flow characteristic of various vocal characteristics in the glottal source pulses.

### 4.2 Modelling Vocal Characteristics and Formant Synthesis

One goal of this study was to demonstrate how a speech synthesizer, specifically a formant synthesizer, can be used to model vocal characteristics caused by laryngeal dysfunction. Reviews of the previous studies on modelling vocal characteristics are given in Lee (1988) and Eskenazi (1988). Most of the previous studies on modelling vocal characteristics involve the analysis of the data from human subjects. However, humans, despite instructions to the contrary, may unknowingly vary one or more parameter (source and vocal-tract characteristics) while phonating a specific sound. Another limitation of such studies is the availability of data from human subjects. Our approach was to develop a unified glottal source model for various vocal characteristics and use it in a formant synthesizer to synthesize speech tokens with various vocal characteristics. The advantage of using a formant synthesizer is that

the source characteristics can be precisely varied independent of the vocal-tract characteristics. Thus, the source characteristics can be precisely controlled and systematically varied to obtain synthesized speech with the desired vocal characteristics. Another advantage is that current implementations of the formant synthesizer are known to produce high quality, natural sounding speech [Childers and Wu, 1990; Klatt and Klatt, 1990; Holmes et al., 1990]. The listeners can perceptually evaluate the naturalness of the vocal characteristics under investigation through the listening tests. Using the formant synthesizer, it is possible to obtain the precise cause-and-effect relationships between the glottal flow characteristics and various vocal characteristics. Lee and Childers (1989) were successful in developing a glottal source model for breathy voice using this approach.

#### 4.3 Glottal Flow Characteristics

The origin of the quasi-periodic pulses of the glottal flow is explained by the myoelastic-aerodynamic theory of vocal-fold vibrations [van den Berg, 1968]. The underlying hypothesis in this study for modeling vocal characteristics is that different vocal characteristics, such as modal, creaky, breathy, rough and hoarse, are represented by distinctive vocal-fold vibratory patterns and thus, by distinctive characteristics of the the volume-velocity of the air flow at the glottis. These glottal flow characteristics can be simulated in the synthesized speech by means of a glottal source model in a formant synthesizer.

Several researchers have analyzed the glottal factors for various vocal characteristics by studying the vibratory movements of the vocal folds and by studying the glottal flow during phonation by normal and pathological subjects. The vocal folds are located below the pharynx making it difficult to observe vocal-fold movements or to measure glottal flow. The methods for indirect observation of the vocal folds

include X-ray laminagraphy [Allen and Hollien, 1973], laryngeal stroboscopy [Hirano, 1981], and ultra-high speed laryngeal cinematography [Childers et al., 1983]. The methods for indirect measurement of glottal flow include electroglottography (EGG) [Childers and Krishnamurthy, 1985], photoglottography (PGG) [Kitzing, 1982], ultra-sound glottography (UGG) [Hamlet, 1981], and inverse filtering of speech signal [Davis, 1976; Lee and Childers, 1989; Eskenazi et al., 1990]. These and several other researchers [Liberman, 1961 and 1963; Koike et al., 1977; Monsen and Engebretson, 1977; Horii, 1980; Yumoto et al., 1982 and 1984; Hiraoka et al., 1984; Lee and Childers, 1989; Pinto and Titze, 1990] have defined several features of the glottal flow that are related to various vocal characteristics. We have listed a few glottal factors that are known to be significant for characterizing various vocal characteristics.

#### 4.3.1 Time Domain Glottal Factors

The time domain glottal factors are based upon the shape of the glottal flow pulses. Some of the time domain glottal factors are:

- 1) Pitch period ( $t_0$ ): time interval between the onset of two glottal flow pulses.
- 2) Glottal flow pulse width: duty cycle of a glottal flow pulse (open quotient, OQ).
- 3) Glottal flow pulse skewness: the ratio of opening phase to the closing phase of a glottal flow pulse (speed quotient, SQ).
- 4) Abruptness of closure (AC): the rate of decrease in the glottal flow at the termination of a glottal flow pulse.
- 5) Aspiration noise: The turbulent noise generated at the glottis as the air is expelled through the incomplete closure of the vocal folds; characterized by the signal-to-noise ratio, SNR.
- 6) Pitch perturbation (Jitter): variation in the pitch period (duration of glottal flow pulse) of the glottal source pulses in a sustained phonation.



7) Amplitude perturbation (shimmer): variation in the peak amplitude of each of the glottal flow pulses in a sustained phonation.

#### 4.3.2 Frequency Domain Glottal Factors

The frequency domain glottal factors are based upon the magnitude frequency response of the waveform of the glottal flow pulses (glottal flow spectrum). Some of the frequency domain glottal factors are

- 1) Fundamental frequency (F0): the rate of generation of glottal flow pulses.
- 2) Spectral tilt (ST): the asymptotic slope of the envelop of the glottal flow spectrum in the high-frequency region (also known as spectral slope).
- 3) Harmonic Richness Factor (HRF): the ratio of the sum of the magnitude of the harmonics to the magnitude of the fundamental frequency component in the glottal flow spectrum [Lee and Childers, 1989].
- 4) Harmonic to Noise Ratio (HNR): the ratio of the power in the harmonics to the power in between the harmonics in the glottal flow spectrum. This measure is based on Harmonic to Noise Ratio (HNR) defined by Hiraoka et al. (1984), Noise to Harmonic Ratio (NHR) defined by Lee and Childers (1989) and Signal to Noise Ratio (SNR) defined by Kojima et al. (1980).

#### 4.4 Vocal Characteristics and Glottal Factors

The significance of glottal factors in characterizing various vocal characteristics has been studied in terms of statistical correlations between one or more of the above listed glottal factors and the vocal characteristics such as, modal, creaky, breathy, rough, hoarse, etc. [Davis, 1976; Wolf and Steinfath, 1987; Yumoto et al., 1982; Eskenazi et al., 1990]. These statistical correlations are based upon the values of the glottal factors obtained from the analysis of speech signals collected from the normal

and pathological subjects and upon the ratings for the severity of each vocal characteristic perceived by trained speech researchers and clinicians. In the following paragraphs we give a summary of the characteristics of the glottal factors for creaky, modal, breathy, rough and hoarse vocal characteristics. The characteristics of the glottal factors for each vocal characteristics are also listed in Table 4-I. This summary and table are based upon the survey conducted by Lee (1988) and Eskenazi (1988). The readers are referred to Lee (1988) and Eskenazi (1988) for a detailed description of glottal flow characteristics for various vocal disorders.

Creaky and modal voices are two vocal registers. A creaky voice is characterized by a low fundamental frequency between 18 to 46 Hz for male speech and 24 to 52 Hz for female speech. The glottal flow pulses for creaky phonations have a short pulse width (about 25% of the pitch period), a long closed interval, may have multiple opening and closing intervals within a pitch period, high pulse skewness and very abrupt closure. The frequency domain characteristics are a small spectral tilt and high harmonic richness factor.

It is difficult to characterize modal or normal voice, since it depends upon several factors, such as, age, sex, emotional state, cultural group, etc. Therefore, the category of the modal voice encompasses a broad range of phonations. Modal voices are characterized by a medium range of fundamental frequency, between 94 to 287 Hz for male and 144 to 538 Hz for female speech. The glottal flow pulses for modal phonations have a medium pulse width (66% of the pitch period), short closed interval, medium pulse skewness and an abrupt closure. Pitch perturbation and amplitude perturbation are present even in the sustained phonations of vowels. A small amount of aspiration noise is also present, particularly, in the slightly breathy sounding female phonations. The frequency domain characteristics are a medium spectral tilt, medium harmonic richness factor and high harmonic to noise ratio.

Table 4-I  
 Characteristics of Glottal Factors for Various Voice Types  
 Time Domain Glottal Factors

| Parameter | Pitch Period ( $t_0$ ) | Pulse Width (OQ) | Pulse Skewness (SQ) | Abruptness of Closure (AC) | Aspiration Noise (SNR) | Jitter | Shimmer |
|-----------|------------------------|------------------|---------------------|----------------------------|------------------------|--------|---------|
| Modal     | Medium                 | Medium           | Medium              | Medium                     | High                   | Low    | Low     |
| Creaky    | Long                   | Short            | High                | Fast                       | Medium                 | High   | High    |
| Breathy   | Medium                 | Long             | Low                 | Slow                       | Low                    | Low    | Medium  |
| Rough     | Medium                 | Medium           | Medium              | Medium                     | Medium                 | High   | High    |
| Hoarse    | Medium                 | Medium           | Medium              | Medium                     | Low                    | High   | High    |

Frequency Domain Glottal Factors

| Parameter | Fundamental Frequency ( $F_0$ ) | Spectral Tilt (ST) | Harmonic Richness Factor (HRF) | Harmonic to Noise Ratio (HNR) |
|-----------|---------------------------------|--------------------|--------------------------------|-------------------------------|
| Modal     | Medium                          | Medium             | Medium                         | High                          |
| Creaky    | Low                             | Low                | High                           | Medium                        |
| Breathy   | Medium                          | High               | Low                            | Medium                        |
| Rough     | Medium                          | Medium             | Medium                         | Low                           |
| Hoarse    | Medium                          | Medium             | Medium                         | Low                           |

Severe breathy, rough and hoarse voices are often considered as symptoms of vocal disorders that may be caused by a laryngeal dysfunction or vocal abuse. The fundamental frequency range for these voices overlap that of modal voice. The glottal flow pulses for breathy phonations have a large pulse width, a very small or absent closed interval, very little pulse skewness and a smooth closure. The main characteristic of a breathy phonation is the presence of high frequency aspiration noise in the glottal flow pulses. The frequency domain characteristics are an approximate spectral tilt of  $-12$  dB/oct or  $-18$  dB/oct, low harmonic richness factor and low harmonic to noise ratio. The glottal flow pulses for the rough phonations have a medium pulse width, medium pulse skewness and an abrupt closure. The main characteristic of rough phonations is the presence of pitch period perturbation. The frequency domain characteristics are an approximate spectral tilt of  $-12$  dB/oct, medium spectral tilt, medium harmonic richness factor and low harmonic to noise ratio. The glottal flow pulses for the hoarse phonations are similar to those for the rough phonations. The main characteristic of the hoarse phonations are the presence of both aspiration noise and pitch period perturbation. The frequency domain characteristics are an approximate spectral tilt of  $-12$  dB/oct, medium harmonic richness factor and very low harmonic to noise ratio.

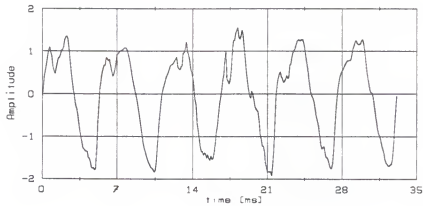
#### 4.5 Glottal Source Model

The objective of this study was to find the relationships between the time and frequency domain glottal factors and various vocal characteristics in order to develop models for various vocal characteristics. Toward that end, we have developed a new unified glottal source model [Lalwani and Childers, 1991b]. This glottal source model is implemented in a formant synthesizer to synthesize speech tokens with various vocal characteristics. Using this model, we can generate glottal source pulses by

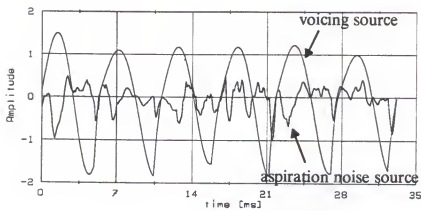
systematically varying glottal flow characteristics. By listening to the speech tokens synthesized from these glottal source pulses, we can find the significance of each glottal factor and also the effect of variation of one or more of these glottal factors on the perception of each vocal characteristic.

A waveform of differentiated glottal flow pulses obtained by inverse filtering of a speech signal with a breathy characteristic is shown in Figure 4-1. The waveforms of differentiated glottal flow pulses for various vocal characteristics are similar in nature. When simulating these waveforms via a glottal source model, the normal practice is to treat such waveforms as being composed of "smooth" signal pulses and the additive random noise. It is hypothesized that the volume-velocity at the glottis (voicing source) due to vibration of the vocal folds can be represented by the "smooth" signal pulses and the turbulent air-flow generated at the glottis (aspiration source) can be represented by the additive random noise. The glottal flow pulses, obtained by integration of the differentiated glottal flow pulses, are quasi-periodic in nature and show a varying peak amplitude for each pulse. This quasi-periodicity is considered as perturbation of the mean period (pitch period) of the glottal flow pulses. The varying peak amplitude of the glottal flow pulses is considered as the perturbation of the mean peak amplitude of the glottal flow pulses. The perturbation of the pitch period (pitch perturbation) and the perturbation of the peak amplitude (amplitude perturbation) of the glottal flow pulses are incorporated in the voicing source pulses. The magnitude frequency response of the differentiated glottal source waveform and the "smooth" differentiated voicing source waveform shown in Figure 4-1 are shown in Figure 4-2.

The block diagram of the new glottal source model is shown in Figure 4-3. The new unified glottal source model consists of four types of source models: 1) voicing source model, 2) aspiration noise source model, 3) pitch perturbation source model and 4) amplitude perturbation source model.

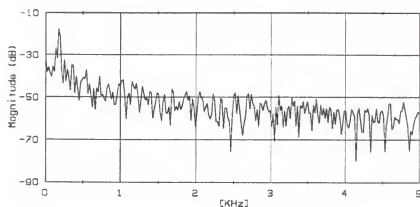


[a]

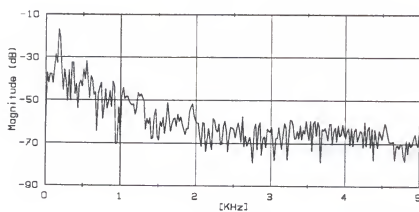


[b]

Figure 4-1: Inverse filtered waveform and its decomposition for modeling  
 a) Inverse filtered waveform of speech with breathy characteristics  
 b) Decomposition into voicing source and aspiration noise source



[a]



[b]

Figure 4-2: Magnitude frequency response  
a) Inverse filtered waveform of speech with breathy characteristics  
b) Corresponding voicing source waveform

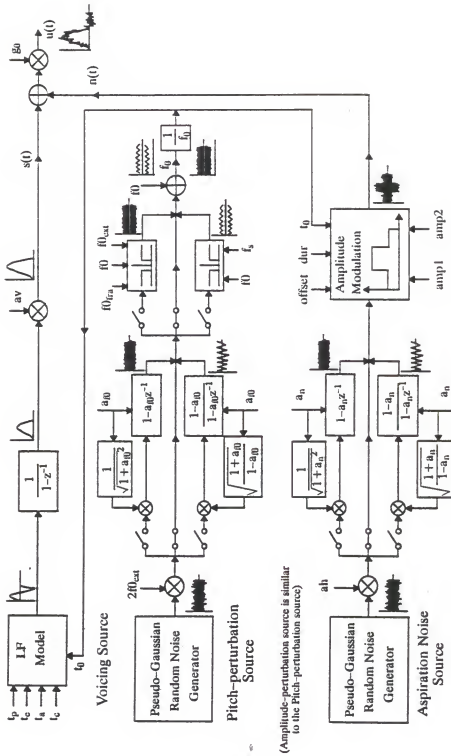


**Figure 4-3: New Glottal Source Model**



#### 4.5.1 Voicing Source Model

A voicing source model generates voicing source pulses that are similar in shape to the glottal source pulses except that they are “smooth.” The glottal factors, such as pitch period of glottal source pulse, glottal source pulse width, glottal source pulse skewness and abruptness of closure of glottal source pulse can be controlled by the voicing source model. These glottal factors determine the overall shape of the voicing source pulses.

We have used the LF (Liljencrants and Fant) model [Fant et al., 1985] as the voicing source model. The time function of the LF model specifies the differentiated voicing source pulse. The voicing source pulse has to be deduced from the integral of the LF time function [Fant et al., 1985; Fant and Lin, 1988]. The LF model generates pulses at the interval specified by the pitch period contour to generate a differentiated voicing source waveform. The LF model consists of three parts:

L model:

$$\frac{dU_g(t)}{dt} = E_0 e^{at} \sin(\omega_g t) \quad 0 < t \leq t_e$$

Recovery phase model:

$$\frac{dU_g(t)}{dt} = \frac{-E_e}{\xi t_a} [e^{-\xi(t-t_e)} - e^{-\xi(t_c-t_e)}] \quad t_e < t \leq t_c$$

and

$$\frac{dU_g(t)}{dt} = 0 \quad t_c < t < t_0$$

$t_p$  : the time instant at which the positive peak occurs in the integrated LF time function

$t_e$  : the time instant at which the negative peak occurs in the LF time function

$t_a$  : the time constant of the recovery phase exponential

$t_c$  : the time instant at which the integrated LF time function becomes equal to zero (after the occurrence of positive peak)

- $t_0$  : pitch period, i.e., the time interval between the onset of two consecutive LF time functions in the glottal source waveform,  $t_p < t_e \leq t_c \leq t_0$
- $E_e$  : absolute value of the negative peak in the LF time function
- $\omega_g$  : frequency of the exponentially increasing sinusoid (the L model)
- $\alpha$  : growth constant of the exponentially increasing sinusoid,  $\alpha > 0$  for exponential increase
- $E_0$  : scale factor for the L model time function to obtain the negative peak equal to  $-E_e$  at time  $t_e$
- $\xi$  : decay constant of the recovery phase exponential,  $\xi > 0$  for exponential decay

It can be observed that the LF model time function is obtained by appending the recovery phase time function to the L model time function at time instant  $t_e$ . A typical LF model time function (differentiated voicing source pulse) and its integral (voicing source pulse) are shown in Figure 4-4.

There are two constraints that have to be satisfied by the L model and recovery phase time functions.

- 1) In order to maintain continuity between the L model and the recovery phase, both the time functions should be equal to  $-E_e$  at time instant  $t_e$ .
- 2) The total area under the time function of the LF model should be zero so that the integral of the LF model time function is zero at time instant  $t_c$ , i.e., the voicing source pulse has zero value at time instant  $t_c$ .

The parameters " $t_p$ ," " $t_e$ ," " $t_a$ " and " $t_c$ " are called the "timing parameters" of the LF model. The pitch period " $t_0$ " is specified as the multiplicative inverse of the fundamental frequency parameter " $f_0$ ." The parameters " $E_0$ ," " $\alpha$ ," " $\omega_g$ " and " $\xi$ " are called the "direct synthesis parameters" of the LF model. The LF model time function is generated using the "direct synthesis parameters." In many research applications, it is easier to specify the "timing parameters" and the maximum negative amplitude

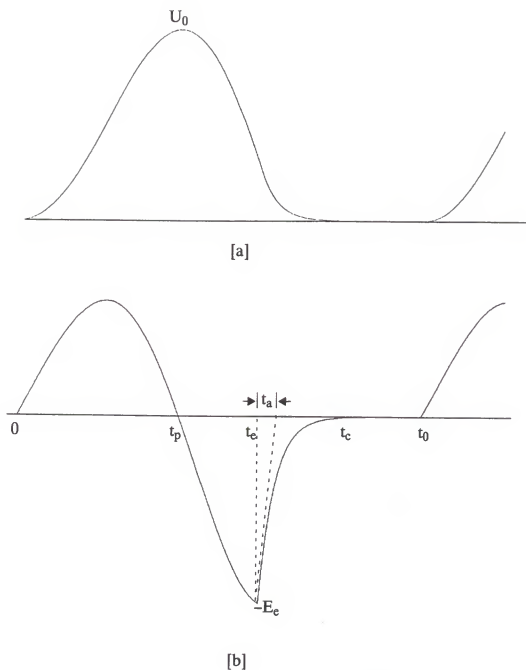


Figure 4-4: LF model time function

- a) Integrated LF model time function (voicing source pulse)
- b) LF model time function (differentiated voicing source pulse)

parameter " $E_e$ " of the LF model than to specify the "direct synthesis parameters" of the LF model [Fant et al., 1985; Gobl, 1988; Gobl, 1989; Karlson, 1986; Lee, 1988]. The values of the "direct synthesis parameters" can be determined from the values of the "timing parameters" by a simple procedure [Fant et al., 1985]. This procedure involves the following steps:

1) Find " $\xi$ " from the parameter " $t_a$ " and the interval  $(t_c - t_e)$ . In order to satisfy the first constraint, the value of the recovery phase curve at time  $t_e$  should be  $-E_e$ . Therefore,

$$\xi t_a = [1 - e^{-\xi(t_c - t_e)}]$$

Solving this nonlinear equation by the Newton-Raphson method we find the value of the parameter " $\xi$ " for the given values of the parameter " $t_a$ " and the interval  $(t_c - t_e)$ .

2) Find the area under the recovery phase time function, with  $E_e = 1$ , by simple integration.

3) Set  $\omega_g = \pi/t_p$ .

4) Use the bisection method to find the value of the parameter " $\alpha$ " such that the area under the LF model time function, with  $E_e = 1$ , is equal to zero or a very small number.

5) The value of the scale factor " $E_0$ " is obtained from the equation

$$E_0 = \frac{-E_e}{e^{\alpha t_e} \sin \omega_g t_e}$$

The advantages of the LF model as a voicing source model are:

1) The LF model is optimal for non-interactive glottal flow parameterization in the sense that it insures an overall fit to commonly encountered glottal flow pulse shapes with a minimum number of parameters and is flexible in its ability to match the glottal flow pulse shapes of extreme phonations [Fant and Lin, 1988; Klatt and Klatt, 1990; Lee, 1988].

2) It can be easily implemented in digital hardware [Fant et al., 1985].

The LF model's parameters specify the differentiated voicing source pulse. The discrete-time differentiated voicing source pulse is obtained by sampling the LF model time function at the sampling rate " $f_s$ ." The voicing source pulse is obtained by integrating the differentiated voicing source pulse using the first order digital IIR filter. The value of the coefficient of this IIR filter is specified by the parameter "gfilt." The value of the parameter "gfilt" should be kept equal to 1.0 to obtain the voicing source pulses that terminate with zero at the closing instant " $t_c$ ." The integrator can be removed from the voicing source by specifying "gfilt" equal to zero, specifically, when the differentiated voicing source pulses are used as excitation source. The parameter "av" specifies the value (in dB) of either the energy, power or the peak amplitude in each voicing source pulse. Which one of the above three quantities the value of the parameter "av" corresponds to, is specified by the parameter "typ\_gain." Accordingly, the voicing source pulses are scaled to make the power, energy or the peak value in each voicing source pulse equal to the value of the voicing gain parameter "av." Each combination of the values of the "timing parameters" result in an unique combination of the values of the glottal factors, such as glottal source pulse width, glottal source pulse skewness and abruptness of closure of a glottal source pulse. The "timing parameters" also determine the values of the frequency domain glottal factors, such as spectral tilt and harmonic richness factor of the glottal source pulses.

#### 4.5.2 Aspiration Noise Source Model

An aspiration noise source model simulates the turbulent air-flow generated at the glottis by an incomplete closure of the vocal folds. The aspiration noise source model can affect the glottal factors, such as signal to noise ratio, harmonic richness factor, harmonic to noise ratio and spectral tilt.

The random number generator in the aspiration noise source model, the pitch perturbation source model and the amplitude perturbation source model each

generate a separate random number sequence (with different seed values). The random number sequences have white-noise characteristics and a pseudo Gaussian distribution. The range of the random number sequence is  $\pm 0.5$  with a zero mean value. The advantage of using a random number sequence with a Gaussian distribution is that if the input to a linear filter is a random number sequence with a Gaussian distribution, the output random number sequence also has a Gaussian distribution. The advantage of using a white-noise source is that it is possible to keep both the average power (mean squared value) and the range (minimum and maximum values of the pseudo-Gaussian distribution) of the output random number sequence the same as the input random number sequence by multiplying the input or the output random number sequence with a scale factor related only to the filter coefficient. The value of the scale factor is given by the multiplicative inverse of the square root of the integral of the square of the impulse response of the filter (from Parseval's theorem).

The random number generator in the aspiration noise source model generates random numbers at the sampling frequency " $f_s$ ." The random numbers are multiplied by a scale factor such that the power in a long random number sequence is unity. The average power of the aspiration noise source (random number sequence) for the duration of the pitch period is set equal to the value specified by the parameter "ah" (in dB) by further modification of the random number sequence with an additional scale factor. Klatt (1980) has used a first order IIR filter in series with the aspiration noise source to simulate the volume-velocity due to a turbulent noise (pressure) source at a constriction in the vocal-tract. Lee and Childers (1989) have used a FIR filter in series with the aspiration noise source to simulate the volume-velocity of the air flow due to turbulent noise source at the glottis in their model for breathy voice. In the new glottal source model the aspiration noise source may be filtered by a first order FIR filter or a first order IIR filter using the FOS (First Order System). The value of the filter coefficient " $a_n$ " specifies the type of filter: a first order FIR filter, a first order

IIR filter or a by-pass path. If the value of the parameter " $a_n$ " is equal to zero, the aspiration noise source is not filtered. If the value of the parameter " $a_n$ " is between 0.0 and 1.0, the aspiration noise source is filtered by a first order IIR (lowpass) filter. If the value of the parameter " $a_n$ " is between -1.0 and 0.0, the aspiration noise source is filtered by a first order FIR (highpass) filter. The bandwidth (in Hz) of the passband of the FIR and the IIR filters is determined by both the sampling frequency, " $f_s$ ," and the filter coefficient, " $a_n$ ." The aspiration noise source is scaled by a factor proportional to the parameter " $a_n$ ," prior to filtering, in order to maintain the average power after filtering the same as that before filtering (discussed at the beginning of this sub-section). In the new glottal source model, the shape of the spectrum of an aspiration noise source can be varied independently of the average power of that aspiration noise source, and vice versa. To illustrate the above feature, a unit power aspiration noise source of 20000 samples was generated and filtered by a lowpass filter and a highpass filter. Table 4-II shows the power of the aspiration noise source, with and without multiplication by the scale factor, for each of the three cases: unfiltered, lowpass filtered and highpass filtered aspiration noise source. The spectrum and distribution of the aspiration noise source for each of these three cases are shown in Figure 4-5 and Figure 4-6.

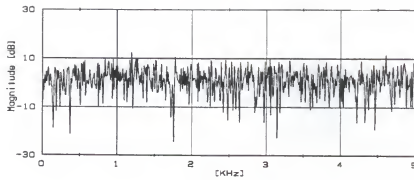
The aspiration noise source may then be amplitude-modulated by an amplitude-time waveform. Amplitude modulation simulates both the effect of vibrating vocal folds on the steady air flow from the lungs when phonating the sounds with mixed-excitation [Klatt, 1980] and also the generation of the turbulent air flow due to incomplete closure of the vocal folds during breathy phonations [Lee and Childers, 1989]. The amplitude-modulation waveform is the same as that described in Chapter 2. The amplitude-modulation waveform has three parts for each glottal source pulse. The parameter " $amp1$ " specifies the amplitude of the first and the third parts and the parameter " $amp2$ " specifies the amplitude of the second part. The

Table 4-II

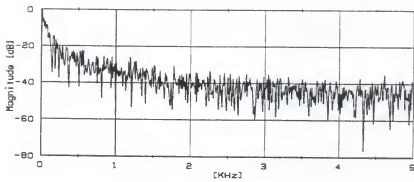
Aspiration noise power with/without scaling

| Type of aspiration noise source | Power without scaling | Power with scaling |
|---------------------------------|-----------------------|--------------------|
| without filtering               | 1.083                 | 1.083              |
| lowpass filtered                | 0.00522               | 1.039              |
| highpass filtered               | 2.103                 | 1.052              |

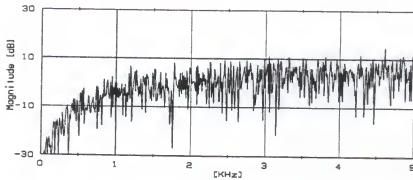




[a]



[b]



[c]

Figure 4-5: Frequency domain characteristics of aspiration noise source  
a) Spectrum of unfiltered aspiration noise source  
b) Spectrum of lowpass filtered aspiration noise source  
c) Spectrum of highpass filtered aspiration noise source

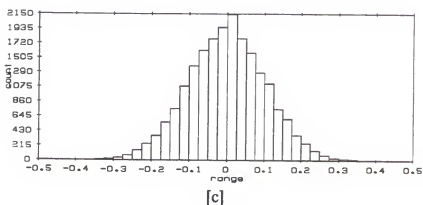
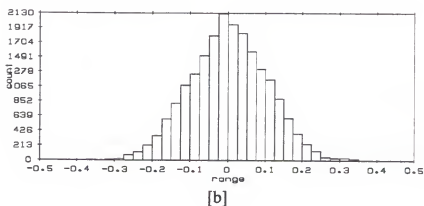
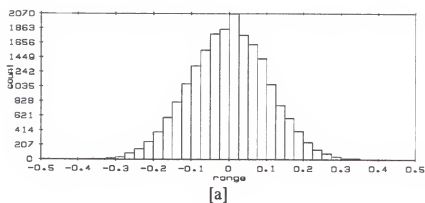


Figure 4-6: Histogram of aspiration noise source  
 a) Unfiltered  
 b) Lowpass filtered (after scaling)  
 c) Highpass filtered (after scaling)

parameters “offset” and “dur” specify the duration of the first and the second parts, respectively. The duration of the third part is given by the value (pitch period – (offset + dur)).

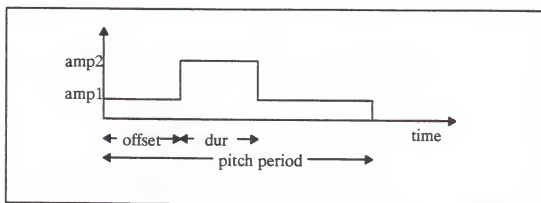
A typical amplitude-modulation waveform is shown in Figure 4-7. Amplitude-modulation of the aspiration noise source by an amplitude-modulation waveform with  $\text{amp1} = 0.0$ ,  $\text{amp2} = 1.0$ ,  $\text{offset} = 50\%$  and  $\text{dur} = 50\%$  and pitch period = 10 msec is also shown in Figure 4-7. After amplitude-modulation, the average power in the aspiration noise source for the duration of pitch period decreases to half of the initial value specified by the “ah” parameter.

The discrete samples of voicing source pulses and aspiration noise source are added together to obtain the discrete glottal source pulses. For each glottal source pulse, the power in the voicing source pulse, specified by the parameter “av,” can be considered as the signal power. The power in the aspiration noise source for the duration of the pitch period is specified by the parameter “ah.” Therefore, the signal to noise ratio (SNR) for each glottal source pulse is

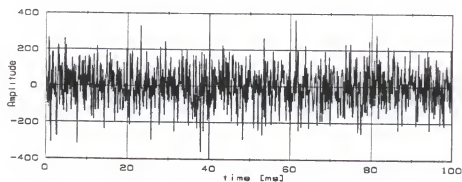
$$\text{SNR} = \text{av} - \text{ah} \text{ (in dB)}$$

The value of SNR for each glottal source pulse can be controlled by specifying the values of “av” and “ah” parameters appropriately. The overall amplitude of the glottal source waveform is controlled by the parameter “g<sub>0</sub>” (in dB).

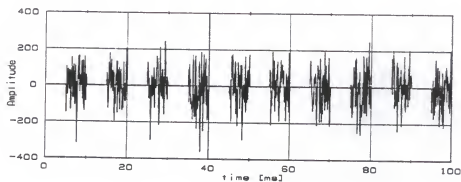
The Figure 4-8 shows: a) voicing source pulse, b) voicing source pulse plus aspiration noise source (SNR = 30 dB) and c) voicing source pulse plus aspiration noise source (SNR = 20 dB). The spectrum of the glottal source pulses in each of these cases is shown in the Figure 4-9. One can observe that the addition of aspiration noise to the voicing source pulses reduces the amplitude of the harmonic components and increases the amplitude of the inter-harmonic components, i.e., causes a decrease in the values of the glottal factors, such as spectral tilt, harmonic richness factor and harmonic to noise ratio.



[a]

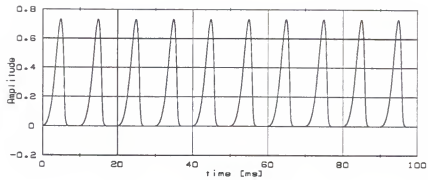


[b]

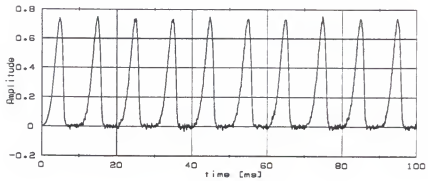


[c]

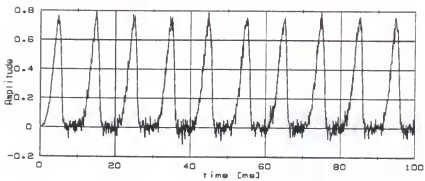
Figure 4-7: Amplitude modulation of noise source  
 a) Amplitude-modulation waveform  
 b) Aspiration noise source  
 c) Amplitude-modulated aspiration noise source  
 with  $\text{amp1} = 0.0$ ,  $\text{amp2} = 1.0$ ,  $\text{offset} = 0.5$  and  $\text{dur} = 0.5$



[a]

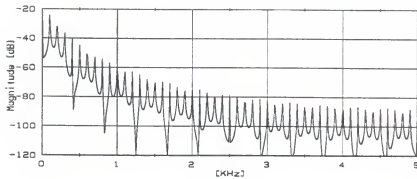


[b]

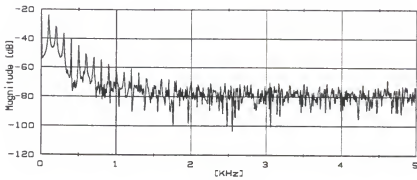


[c]

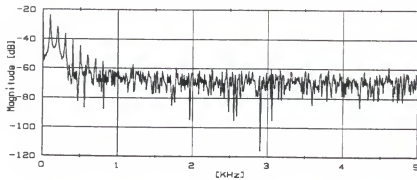
Figure 4-8: Glottal source pulses  
a) Only voicing source  
b) Voicing source plus aspiration noise source (SNR 30 dB)  
c) Voicing source plus aspiration noise source (SNR 20 dB)



[a]



[b]



[c]

Figure 4-9: Spectra of glottal source pulses shown in the Figure 4-8  
 a) Only voicing source  
 b) Voicing source plus aspiration noise source (SNR 30 dB)  
 c) Voicing source plus aspiration noise source (SNR 20 dB)

#### 4.5.3 Pitch Perturbation Source Model

A pitch perturbation source model simulates the period-to-period variations in the fundamental frequency parameter, “ $f_0$ ,” (equivalently the pitch period, “ $t_0$ ” of the glottal source pulses). The variation in the fundamental frequency of the glottal source pulses incorporates the perceptual equivalent of “jitter” in the synthesized speech. The pitch perturbation source model can control such glottal factors as harmonic richness factor and harmonic to noise ratio.

The random number generator in the pitch perturbation source generates random numbers at the “frame rate.” The pitch perturbation sequence has a pseudo-Gaussian distribution within the range  $\pm 0.5$  and a zero mean value. Pinto and Titze (1990) have shown that the pitch perturbations in the natural speech may not necessarily have a Gaussian distribution. Kobayashi and Sekine (1990) have observed a gamma distribution of the pitch perturbation in the natural speech. But they observed no significant difference in the listeners response, whether the distribution of the pitch perturbations was uniform, Gaussian or the gamma distribution in the synthesized speech tokens. In our model, the “extent of the pitch perturbation” of a pitch perturbation sequence is specified as a fraction of the mean fundamental frequency “ $f_{0\text{mean}}$ ” by the parameter “ $f_{0\text{ext}}$ .” The parameter “ $f_0$ ” specifies the mean fundamental frequency, “ $f_{0\text{mean}}$ .” The random number sequence is multiplied by a scale factor equal to  $2*f_{0\text{ext}}*f_{0\text{mean}}$ . After multiplying by the scale factor, the pitch perturbation sequence has pseudo Gaussian distribution within the range  $\pm f_{0\text{ext}}*f_{0\text{mean}}$  and a zero mean value. The reason for choosing the scale factor equal to  $2*f_{0\text{ext}}*f_{0\text{mean}}$  is as follows:

- 1) The maximum absolute value of the pitch perturbation in the pitch perturbation sequence is  $f_{0\text{ext}}*f_{0\text{mean}}$ . It is proportional only to the parameter “ $f_{0\text{ext}}$ ” when the

fundamental frequency is specified as constant (which is normally the case when synthesizing sustained phonations).

2) The pitch perturbation is proportional to the mean fundamental frequency, and hence, the values of the “Jitter Factor” and the “Frequency Perturbation Quotient” measures (described later) in the synthesized speech are independent of the mean fundamental frequency.

Pinto and Titze (1980) have used the zero crossing rate (ZCR) as a measure of the rate of change of pitch perturbation in a pitch perturbation sequence. In our glottal source model, we control the rate of change of pitch perturbation by changing the spectrum of the pitch perturbation sequence using either a first order FIR (highpass) filter or a first order IIR (lowpass) filter. When the pitch perturbation sequence is highpass filtered, the values of the pitch perturbation change rapidly. As the bandwidth of the passband of a highpass filter increases the rate of change of the pitch perturbation decreases. When the pitch perturbation sequence is lowpass filtered, the values of the pitch perturbation change slowly. As the bandwidth of the passband of a lowpass filter increases the rate of change of the pitch perturbation increases. AFOS is used to provide options for selecting the type of filter for varying the “zero crossing rate of the pitch perturbation.” The bandwidth (in Hz) of the passband of the FIR and the IIR filter is determined by both the “frame rate” (in Hz) and the filter coefficient “ $a_0$ .” The pitch perturbation sequence is multiplied by a scale factor proportional to the parameter “ $a_0$ ” prior to filtering to maintain the same “extent of the pitch perturbation” in the pitch perturbation sequence after filtering the same as that before the filtering (as discussed at the beginning of the previous sub-section). In the new glottal source model, the “extent of the pitch perturbation” of a pitch perturbation sequence can be varied independently of the “rate of pitch perturbation.” To illustrate these features, a pitch perturbation sequence with 200 samples was generated and filtered separately by a lowpass filter and a highpass filter. The



histogram of the pitch perturbation sequences for the three cases: unfiltered, lowpass filtered and highpass filtered pitch perturbation sequence are shown in the Figure 4–10.

We have chosen to add pitch perturbation to the fundamental frequency parameter, “ $f_0$ ,” instead of the pitch period. The parameter “ $f_0$ ” is an independent parameter where as pitch period depends upon both the “ $f_0$ ” and “ $f_s$ ” parameters. In a discrete system, the pitch period is determined by rounding-off the ratio of the “ $f_s$ ” and “ $f_0$ ” parameters to an integer number. Due to round-off, small perturbations to the parameter “ $f_0$ ” may not result in the perturbation of pitch period. This is an implementation problem and is not related to the synthesis of speech with various vocal characteristics. To solve this problem, in the new glottal source model, the pitch perturbation sequence is filtered through a nonlinear function that discards the small nonzero pitch perturbations whose values lie in the cutoff range of that function. Two methods are used to determine the size of the cutoff range of the nonlinear function:

- 1) The values of the lower and upper limits of the cutoff range are set equal to the minimum changes in the value of the fundamental frequency required to decrease and increase the value of the pitch period (integer value of the ratio of “ $f_s$ ” and “ $f_0$ ” parameters) by at least one sample, respectively. This cutoff range is non-uniform; smaller on the lower side and larger on the higher side of the value of the parameter “ $f_0$ .” The cutoff range is proportional to the parameter “ $f_0$ ”; the higher the value of “ $f_0$ ,” the larger is the cutoff range. Therefore, this method is suitable when the value of the parameter “ $f_0$ ” has small values (less than 100 Hz). The cutoff range is independent of the “extent of pitch perturbation,” and therefore, care should be taken when specifying the “extent of pitch perturbation.” The specified “extent of pitch perturbation” may be smaller than the cutoff range at higher frequencies.

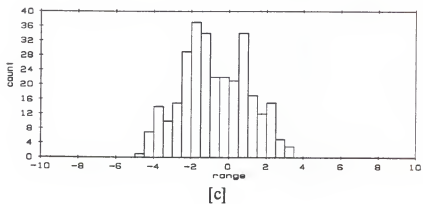
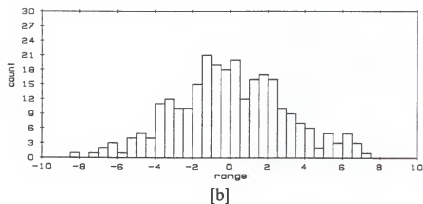
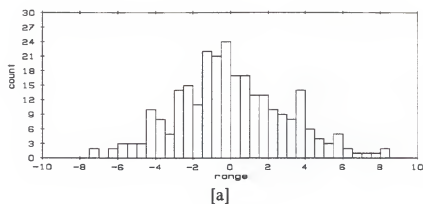


Figure 4-10: Histogram of a pitch perturbation sequence  
 a) Without filtering  
 b) Highpass filtered (after scaling)  
 c) Lowpass filtered (after scaling)

2) The values of the lower and upper limits of the cutoff range are equal to the positive and negative fraction of the “extent of the pitch perturbation,” respectively. This fraction is specified by the parameter “ $f0_{fra}$ .” The cutoff range is uniform and is given by  $\pm f0_{fra} * f0_{ext} * f0$  Hz. Care should be taken while specifying the value of the parameter “ $f0_{fra}$ .” A very small value of “ $f0_{fra}$ ” parameter may result in a small cutoff range which may not be adequate to filter small pitch perturbations.

The unfiltered pitch perturbation sequence, the filtered pitch perturbation sequence and the filtered pitch perturbation sequence passed through a non-linear function for each type of filter and each type of non-linear function are shown in the Figure 4-11, Figure 4-12, Figure 4-13 and the Figure 4-14. Figure 4-15 shows the distribution of the pitch perturbation sequence passed through each of the non-linear functions to illustrate the cutoff range. Note that the non-linear function discards pitch perturbation samples falling in the cut-off region of the non-linear function. Hence, the total number of samples at the output of the non-linear function is less than the original or the filtered pitch perturbation sequence.

After passing the pitch perturbation sequence through a non-linear function, it is added to the fundamental frequency parameter, “ $f0$ ,” to obtain the “fundamental frequency contour.” The value of the pitch period of the voicing source pulse and the amplitude-modulation waveform, for each glottal source pulse, is based upon the value of the parameter “ $f0$ ” in the “fundamental frequency contour.” Figure 4-16 shows the voicing source pulses generated when the value of the parameter “ $f0_{ext}$ ” was 25% and 50% of the fundamental frequency parameter “ $f0$ .” It can be observed that perturbation of the fundamental frequency parameter results in generation of non-periodic voicing source pulses. The spectrum of each of these voicing source pulses is shown in the Figure 4-17. These spectra show reduced amplitude of the harmonic components and increased amplitude of the inter-harmonic components, i.e., a decrease in the values of the glottal factors, such as harmonic richness factor

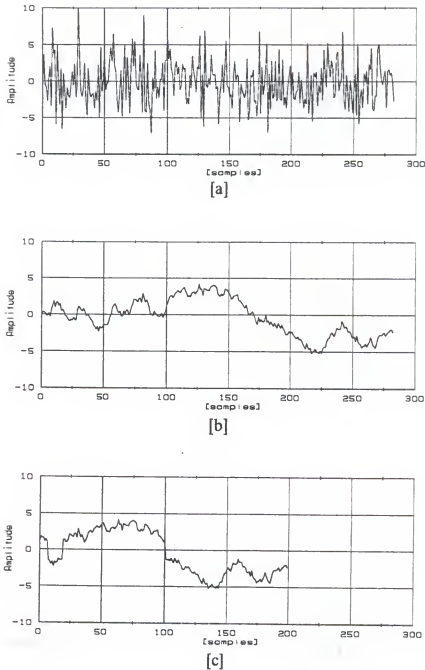


Figure 4-11: Pitch perturbation sequence through lowpass filter and first nonlinearity  
 a) Pitch perturbation sequence with specified range  
 b) Pitch perturbation sequence after lowpass filtering  
 c) Pitch perturbation sequence after lowpass filtering and passing through first nonlinearity. Note that the nonlinear function discards pitch perturbation samples falling in the cut-off region of the nonlinear function. Hence, the total number of samples at the output of the nonlinear function is less than the original or the filtered pitch perturbation sequence.

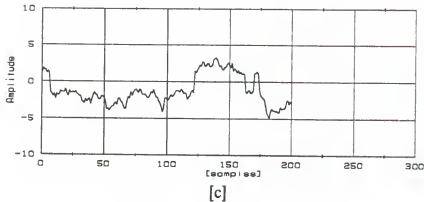
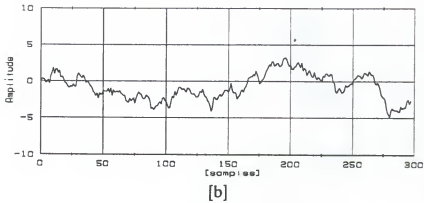
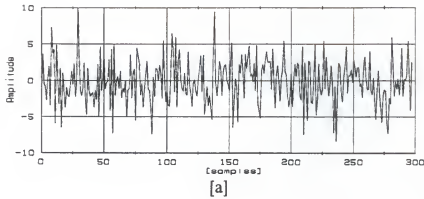


Figure 4-12: Pitch perturbation sequence through lowpass filter and second nonlinearity  
 a) Pitch perturbation sequence with specified range  
 b) Pitch perturbation sequence after lowpass filtering  
 c) Pitch perturbation sequence after lowpass filtering and passing through second nonlinearity.

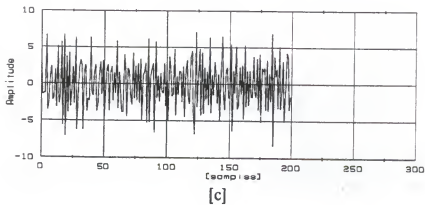
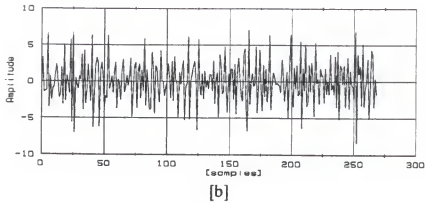
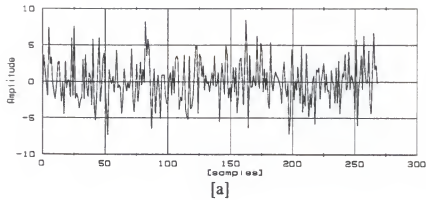


Figure 4-13: Pitch perturbation sequence through lowpass filter and first nonlinearity  
 a) Pitch perturbation sequence with specified range  
 b) Pitch perturbation sequence after highpass filtering  
 c) Pitch perturbation sequence after highpass filtering and passing through first nonlinearity.

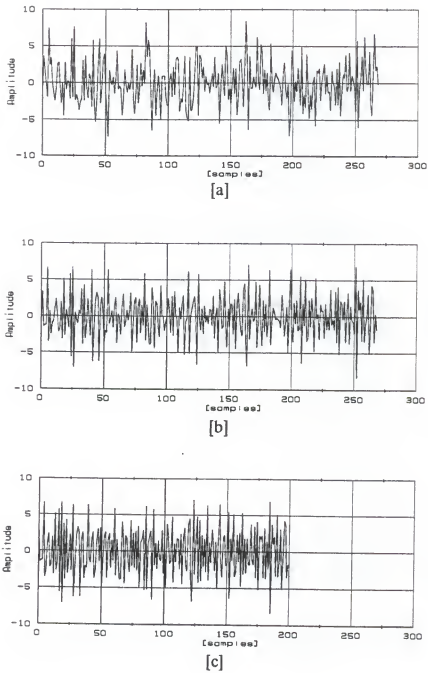
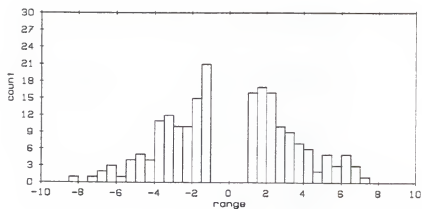
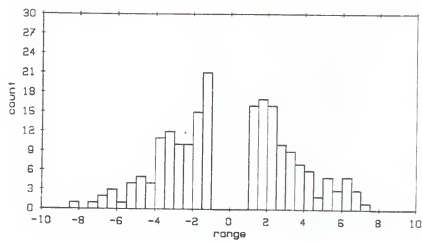


Figure 4-14: Pitch perturbation sequence through highpass filter and second nonlinearity  
 a) Pitch perturbation sequence with specified range  
 b) Pitch perturbation sequence after highpass filtering  
 c) Pitch perturbation sequence after highpass filtering and passing through second nonlinearity.



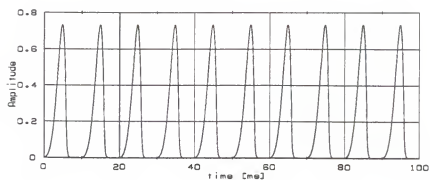
[a]



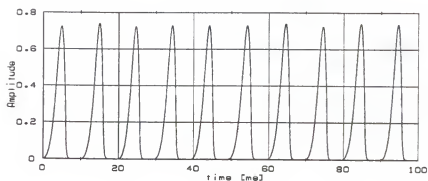
[b]

Figure 4-15: Histogram of pitch perturbation sequence  
a) Highpass filtered and passes through first nonlinearity  
b) Highpass filtered and passes through second nonlinearity

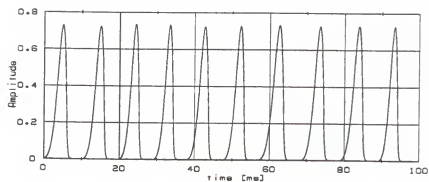




[a]



[b]

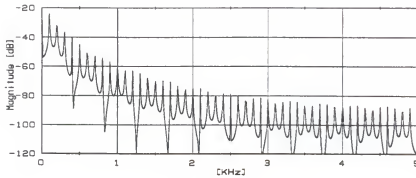


[c]

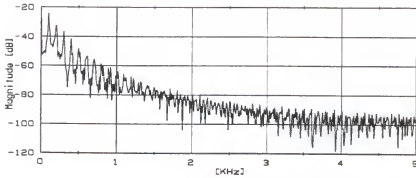
Figure 4-16: Voicing source pulses

a) Only voicing source

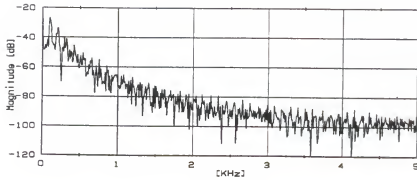
b) Voicing source with pitch perturbation ( $f0_{ext} = 25\%$ )c) Voicing source with pitch perturbation ( $f0_{ext} = 50\%$ )



[a]



[b]



[c]

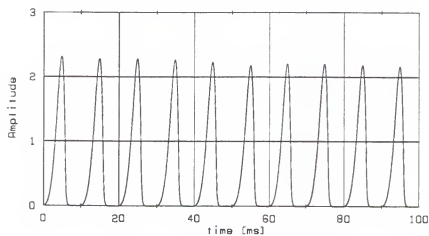
Figure 4-17: Spectra of voicing source pulses in the Figure 4-16  
 a) Only voicing source  
 b) Voicing source with pitch perturbation ( $f0_{ext} = 25\%$ )  
 c) Voicing source with pitch perturbation ( $f0_{ext} = 50\%$ )

and harmonic to noise ratio. It can be observed that in contrast with the spectra shown in Figure 4-9, the level in the high-frequency region in each of the spectra is not increased. Several other measures, such as “Jitter Factor,” “Frequency Perturbation Quotient” and “Directional Jitter” can also be controlled by the pitch perturbation source model. The procedure to control these pitch perturbation measures is described in Appendix D.

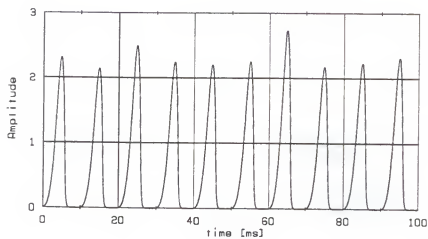
#### 4.5.4 Amplitude Perturbation Source Model

The amplitude perturbation source model is similar to the pitch perturbation source model. The parameter “ $av_{ext}$ ” specifies the “extent of amplitude perturbation,” the parameter “ $a_{av}$ ” determines the “rate of amplitude perturbation” and the parameter “ $av_{fra}$ ” specifies the cutoff range as a fraction of “extent of amplitude perturbation.” The amplitude perturbation sequence is added to the voicing gain parameter “ $av$ ,” to obtain a “gain contour.” The parameter “ $typ\_gain$ ” indicates the quantity (either power, energy or peak amplitude) specified by the “gain contour.” The period-to-period variations in the “gain contour” incorporates the perceptual equivalent of “shimmer” in the synthesized speech.

In the Figure 4-18 the voicing source pulses when the amplitude perturbation sequence is highpass filtered ( $a_{av} = -1.0$ ) and when it is lowpass filtered ( $a_{av} = 0.99$ ) is shown. It can be observed that the addition of perturbation to the voicing gain parameter during synthesis results in generation of non-periodic glottal source pulses. The spectra of the glottal source pulses when the value of the parameter “ $av_{ext}$ ” are 20% and 40 % of the parameter “ $av$ ” are shown in the Figure 4-19. These spectra show a reduced amplitude of the harmonic components and an increased amplitude of the inter-harmonic components, i.e., decrease in the values of the glottal factors, such as harmonic richness factor and harmonic to noise ratio. It can be observed that

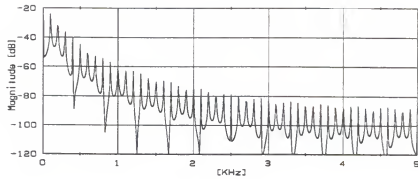


[a]

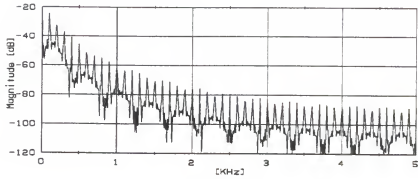


[b]

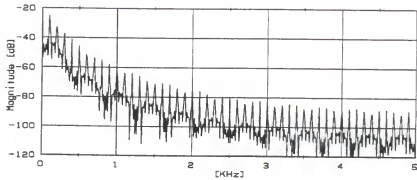
Figure 4-18: Variation in the peak amplitude of the voicing source pulses  
a) amplitude perturbation source lowpass filtered  
b) amplitude perturbation source highpass filtered



[a]



[b]



[c]

Figure 4-19: Spectra of voicing source pulses

a) Only voicing source

a) amplitude perturbation ( $av_{ext} = 20\%$ )b) amplitude perturbation ( $av_{ext} = 40\%$ )

similar to the spectra shown in the Figure 4-17, the level in high frequency region of these three spectra remain approximately the same.

#### 4.6 Summary

Our approach to modeling vocal characteristics is via formant synthesis. We have developed a new unified glottal source model which is implemented in the flexible formant synthesizer to model various vocal characteristics. This model has all the features necessary to generate the glottal source pulses that are typical of various vocal characteristics. The voicing source model can generate voicing source pulses with various shapes to match the glottal flow waveforms commonly observed in typical as well as extreme phonations. A random number generator generates sequences with pseudo Gaussian white-noise characteristics to simulate turbulent air-flow generated at the glottis. Similar random number generators are used to simulate the perturbation of pitch period and amplitude of the glottal flow pulses. In the next chapter we show how various glottal factors (both the time domain and frequency domain glottal factors) in the synthetic speech can be controlled using the parameters of this model.

## CHAPTER 5 GLOTTAL SOURCE MODEL AND GLOTTAL FACTORS

### 5.1 Introduction

In the previous chapter we have described the new glottal source model. We developed this model in order to control various glottal factors known to be significant for modeling various vocal characteristics. This chapter describes the procedures to precisely control both the time domain and frequency domain glottal factors in synthetic speech by appropriate specification of the new glottal source model's parameters.

### 5.2 Controlling Time Domain Glottal Factors

Each time domain glottal factor described in Table 4-I can be directly specified through the new glottal source model's parameters. Also, each time domain glottal factor can be controlled independently of other time domain glottal factors.

#### 5.2.1 Controlling the Pitch Period ( $t_0$ )

The pitch period, " $t_0$ ," of the voicing source pulse is obtained as the multiplicative inverse of the fundamental frequency of the glottal source pulses, which is specified by the parameter " $f_0$ ." The pitch period determines the interval between the onset of voicing source pulses. The change in the rate of repetition of glottal source pulses with the change in fundamental frequency parameter is shown in Figure 5-1.

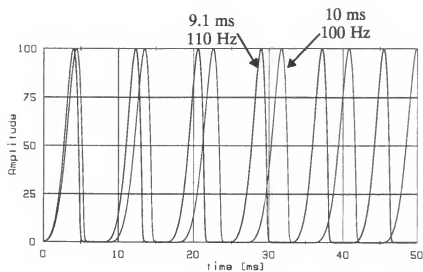


Figure 5-1: Variation in pitch period  
( $t_0$  equal to 10 ms ( $f_0=100\text{Hz}$ )  
and 9.1 ms ( $f_0=110\text{Hz}$ ))



### 5.2.2 Controlling Glottal Source Pulse Skewness (Speed Quotient, SQ)

The glottal source pulse skewness is determined by the skewness of the voicing source pulse, which is determined by the ratio of the duration of the opening phase to the duration of the closing phase of the voicing source pulse. This ratio is also defined as the “speed quotient” (SQ) of the voicing (glottal) source pulse. In the LF model, the duration of the opening phase of the voicing source pulse is specified by the parameter “ $t_p$ .” The duration of the closing phase of the voicing source pulse depends upon the parameters “ $t_e$ ,” “ $t_a$ ,” “ $t_c$ ” and “ $t_0$ .”

If  $t_c \neq t_0$ ,  $SQ = t_p/(t_c - t_p)$ .

If  $t_c = t_0$ ,  $SQ = t_p/(t_e + k \cdot t_a)$ , where  $2 < k < 4$  depending upon  $t_a$ .

The variation in the shape of the differentiated voicing source pulse and the voicing source pulse due to variation in the parameter “ $t_p$ ” ( $t_a = 0$ ,  $t_e = \text{constant}$ ,  $t_c = t_0 = \text{constant}$ ) is shown in Figure 5-2.

### 5.2.3 Controlling the Glottal Source Pulse Width (Open Quotient, OQ)

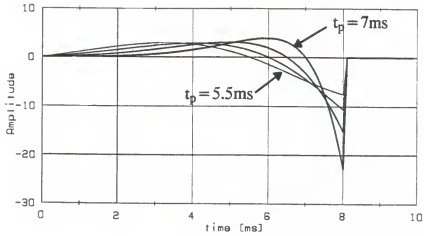
The glottal source pulse width is controlled by the width of the voicing source pulse. The voicing source pulse width is related to the “open quotient” (OQ) of a voicing source pulse, which is defined as the duty-cycle of a voicing source pulse. If the LF model is used as a voicing source model, OQ of the voicing source pulse depends upon the parameters “ $t_e$ ,” “ $t_a$ ,” “ $t_c$ ” and “ $t_0$ .”

If  $t_c \neq t_0$ ,  $OQ = t_c/t_0$ .

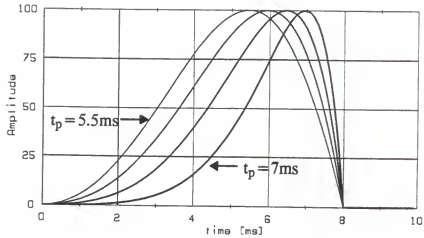
If  $t_c = t_0$  and  $t_a = 0$ ,  $OQ = t_e/t_0$ .

If  $t_c = t_0$  and  $t_a \neq 0$ ,  $OQ = (t_e + k \cdot t_a)/t_0$ , where  $2 < k < 4$  depending upon  $t_a$ .

The variation in the shape of the differentiated voicing source pulse and the voicing source pulse due to variation in the parameter “ $t_e$ ” ( $t_a = 0$ ,  $t_p = \text{constant}$ ,  $t_c = t_0 = \text{constant}$ ) is shown in Figure 5-3. However, we can observe that varying the



[a]



[b]

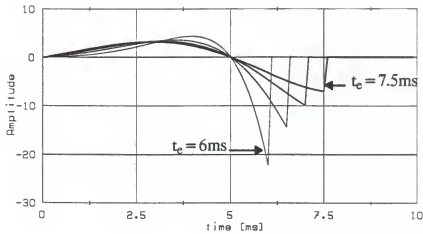
Figure 5-2: Variation in the skewness of the voicing source pulse ( $t_e$  constant at 8 ms and  $t_p$  varying from 5.5 ms to 7 ms in steps of 0.5 ms)

- Differentiated voicing source pulses
- Voicing source pulses

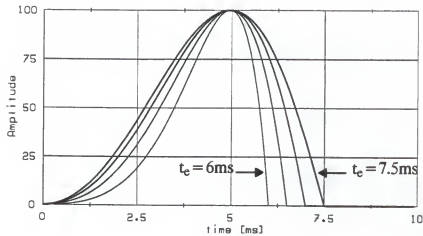
parameter " $t_e$ " to vary OQ also results in a variation of SQ. The variation of OQ while keeping SQ constant requires variation of both the " $t_p$ " and " $t_e$ " parameters. The variation in the shape of the differentiated voicing source pulse and the voicing source pulse when OQ is varied while SQ is kept constant ( $t_a=0$ ,  $t_c=t_0=\text{constant}$ ,  $t_p/(t_e-t_p)=2.0$ ) is shown in Figure 5-4. The parameters " $t_p$ ," " $t_e$ " and " $t_c$ " are specified as the fractions of the pitch period " $t_0$ ," and therefore, variation in fundamental frequency also results in variation of absolute values of these parameters. However, the definition of the glottal factors OQ and SQ of the voicing source pulse involve ratios of the " $t_p$ " " $t_e$ " and " $t_c$ " parameters, and therefore, are independent of the variation in the fundamental frequency.

#### 5.2.4 Controlling the Abruptness of Closure of the Glottal Source Pulse (AC)

The abruptness of closure of the glottal source pulse can be measured as the time constant of the exponential decrease in the glottal flow. The abruptness of closure of a glottal source pulse is controlled by the abruptness of closure of the voicing source pulse. If the LF model is used as a voicing source model, the time constant of the recovery phase time function, " $t_a$ ," controls the abruptness of closure (AC) of the glottal source pulse. If " $t_p$ " and " $t_e$ " are constant and  $t_a=0$ , there is no recovery phase in the voicing source pulse and the abruptness of closure is the maximum for the given values of  $t_p$  and  $t_e$ . As the value of the parameter " $t_a$ " increases the value of the glottal factor AC decreases. The variation in the shape of the differentiated voicing source pulse and the voicing source pulse due to variation in the parameter " $t_a$ " ( $t_p=\text{constant}$ ,  $t_e=\text{constant}$ ,  $t_c=t_0=\text{constant}$ ) is shown in Figure 5-5.

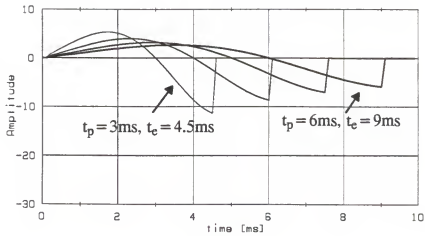


[a]

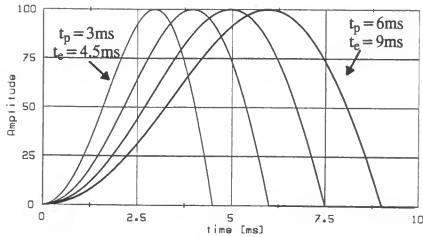


[b]

Figure 5-3: Variation in the width of the voicing source pulse ( $t_p$  constant at 5 ms and  $t_e$  varying from 6 ms to 7.5 ms in steps of 0.5 ms)  
 a) Differentiated voicing source pulses  
 b) Voicing source pulses



[a]



[b]

Figure 5-4: Variation in the voicing source pulse width  
 (when the source pulse skewness is constant)  
 (the ratio  $t_p/(t_c - t_p)$  kept constant at 2, both  $t_p$  and  $t_c$  are variable)  
 a) Differentiated voicing source pulses  
 b) Voicing source pulses

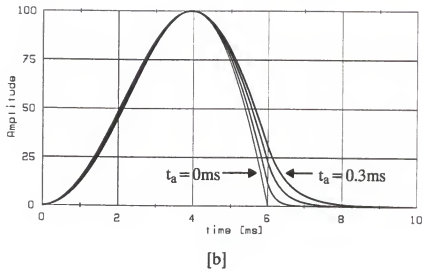
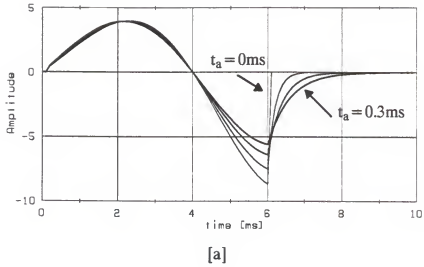


Figure 5-5: Variation in the abruptness of closure of the voicing source pulse ( $t_p$  and  $t_c$  kept constant at 4 and 6 ms, respectively, and  $t_a$  varying from 0 ms to 0.3 ms in steps of 0.1 ms)  
 a) Differentiated voicing source pulses  
 b) Voicing source pulses

### 5.2.5 Controlling the Aspiration Noise (Signal to Noise, Ratio SNR)

The average power in the aspiration noise source for the duration of a pitch period is controlled by the parameter “ah.” The average power in the voicing source pulse for the duration of a pitch period is controlled by the parameter “av.” The ratio of the signal power to the noise power determines the Signal to Noise Ratio (SNR) for each glottal source pulse. The effect of varying SNR on the glottal source pulses and magnitude frequency response is shown in Figure 4–8 and in Figure 4–9.

The power in the aspiration noise source within a pitch period can be further changed by amplitude-modulation. The spectrum of the aspiration noise source is controlled by the filter coefficient “am.” The effect of changes in the amplitude-modulation and/or the spectrum of the aspiration noise source cannot be easily illustrated in figures. One has to listen to synthesized speech in order to evaluate the effect of these changes.

### 5.2.6 Controlling “Jitter” Using the New Glottal Source Model

From a review of the literature on vocal disorders it was determined that “jitter,” i.e., the period-to-period variation in the fundamental frequency parameter, can be measured with the Jitter Factor (JF) [Hollien, 1975], Frequency Perturbation Quotient (FPQ) [Koike et al., 1977] and the Directional Jitter (DJ) [Hecker and Kreul, 1971]. These measures are defined in Appendix D. These measures are significant for differentiating between the “normal” voice and “pathologic” voice. These measures have higher values for “pathologic” voice than for the “normal” voice. Using the new glottal source model, it is possible to synthesize a speech signal with desired values of JF, FPQ and DJ measures. Also, we have incorporated features in the new glottal source model to change the JF and FPQ measures independent of DJ measure, and conversely, the DJ measures independently of the JF and FPQ measures.

The JF and FPQ measures can be controlled by the parameter, “ $f0_{ext}$ ,” which specifies the desired “extent of pitch perturbation,” i.e., the range of the values of pitch perturbation, as a fraction of the mean value of the fundamental frequency parameter. The higher the value of the “ $f0_{ext}$ ” parameter, the higher are the values of the JF and FPQ measures. The analytical expressions for these measures have been derived in Appendix D. The analytical values of JF and FPQ measures in terms of “ $f0_{ext}$ ” parameter are given by

$$JF = 100 * \sqrt{2} * 1.6 * f0_{ext} * \sigma_{md}$$

and

$$FPQ = \frac{\sqrt{6} * 1.6 * f0_{ext} * \sigma_{md}}{3}$$

( $\sigma_{rnd}$  is normally constant for a random number generator)

The DJ measure can be controlled by the “rate of pitch perturbation” parameter, “ $a_0$ ,” which specifies the value of the coefficient of a first order FIR or a first order IIR filter used for changing the zero crossing rate of change of pitch perturbation of the pitch perturbation sequence. The relationship between the parameter, “ $a_0$ ,” the filter type, bandwidth of the passband of the lowpass and highpass filter and the DJ measure is illustrated in Figure 5–6 and in Appendix D. It is observed that the DJ measure increases as the bandwidth of the highpass filter decreases. The DJ measure decreases as the bandwidth of the lowpass filter decreases.

In the new glottal source model we have made provision to keep either the JF or the FPQ measure constant while the DJ measure is varied. The parameter “jmeas\_typ” is used to select one of the two pitch perturbation measure to be kept constant. If selected, the value of the JF measure is kept constant at  $\sqrt{2} * 1.6 * f0_{ext} * \sigma_{rnd}$ . If selected, the value of the FPQ measure is kept constant at  $\sqrt{6} * 1.6 * f0_{ext} * \sigma_{rnd}$ . Only the selected pitch perturbation measure can be kept constant (i.e., will be independent of filter type and the filter coefficient); the other pitch perturbation measure will vary.



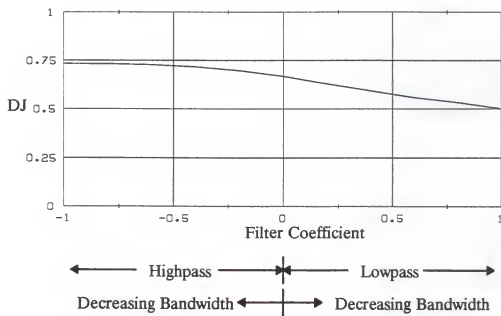


Figure 5-6: Directional Jitter (DJ) versus filter coefficient

### 5.2.7 Controlling “Shimmer” Using the New Glottal Source Model

“Shimmer” is defined as the perturbation of the peak amplitude in each of the glottal source pulses. “Shimmer” can be measured as the mean of the absolute value of the fluctuations in the peak amplitude of the glottal source pulse.

$$Shimmer = \frac{1}{N-1} \sum_{i=2}^N |av_i - av_{i-1}|$$

where the parameter “av” specifies the peak amplitude of a glottal source pulse. In the new glottal source model, the parameter “av<sub>ext</sub>” specifies the “extent of amplitude perturbation,” the parameter “a<sub>av</sub>” specifies the “rate of amplitude perturbation.” The amplitude perturbation source model is similar to the pitch perturbation model, and therefore, the parameters “av<sub>ext</sub>” and “a<sub>av</sub>” can be used to control amplitude perturbation in synthetic speech, similar to the manner in which the parameters “f0<sub>ext</sub>” and “a<sub>f0</sub>” are used for controlling pitch perturbation in synthetic speech.

### 5.3 Controlling Frequency Domain Glottal Factors

The frequency domain glottal factors described in Table 4–I are related to the time domain glottal factors described in the same table. These frequency domain glottal factors are correlated, and hence, cannot be independently controlled. Also, the frequency domain glottal factors cannot be directly specified through the new glottal source model’s parameters. Therefore, we find the values of such frequency domain glottal factors for different combinations of the related time domain glottal factors. We illustrate the relationships between the frequency domain and the related time domain glottal factors in the form of graphs. From these graphs one can obtain the value of a frequency domain glottal factor for a given combination of time domain glottal factors, or, find the values of time domain glottal factors required to achieve a desired value of a frequency domain glottal factor.

### 5.3.1 Controlling the Fundamental Frequency (F0)

The fundamental frequency of the glottal source pulses,  $F_0$ , is determined by the fundamental frequency of the voicing source pulses, which is specified by the parameter "f0." The higher the value of fundamental frequency parameter, the farther apart and fewer are the harmonics in the magnitude frequency response of the voicing source pulses as shown in Figure 5-7.

### 5.3.2 Controlling the Spectral Tilt

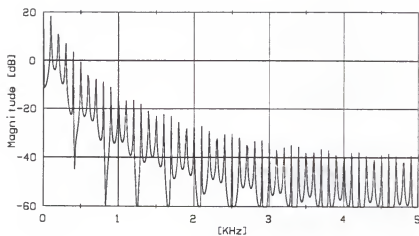
The glottal factor spectral tilt (ST) is defined as the asymptotic slope of the envelope of the magnitude frequency response of the glottal source pulses at high frequencies. The level of the magnitude frequency response of a glottal source pulse at high frequencies depends upon the level of the envelope of the high frequency harmonics, aspiration noise spectra and the extent of the pitch perturbation. It is assumed that the spectral tilt, ST, is determined by the asymptotic slope of the envelope of the high frequency harmonics. The actual level of the magnitude frequency response may be different from the one determined by the glottal factor ST because of the change in the amplitude of the harmonic and inter-harmonic components caused by additive noise and pitch perturbation. This assumption is illustrated in Figure 4-9 and Figure 4-17.

The amplitude of the high frequency harmonics depends upon the abruptness of closure of the voicing source pulses. When the LF model is used as a voicing source model, ST can either be -12 dB/oct or -18 dB/oct depending upon the parameter  $t_a$ .

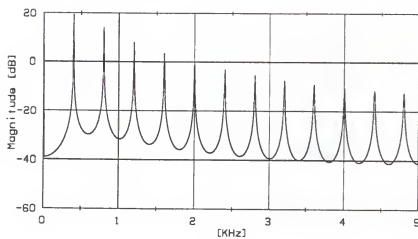
If  $t_a = 0$ ,  $ST = -12 \text{ dB/oct}$ .

If  $t_a \neq 0$ ,  $ST = -18 \text{ dB/oct}$ .

The magnitude frequency response and the spectral tilt of the differentiated LF model time function when  $t_a = 0$  and when  $t_a \neq 0$  are shown in Figure 5-8.

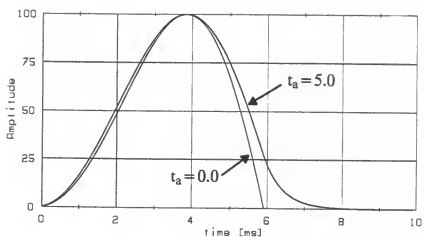


[a]

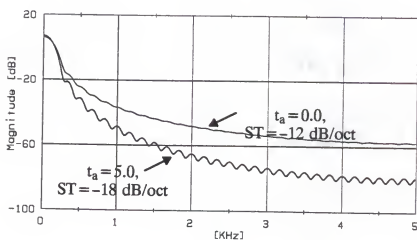


[b]

Figure 5-7: Magnitude frequency response of the voicing source pulses  
a)  $F_0 = 100$  Hz  
b)  $F_0 = 400$  Hz



[a]



[b]

Figure 5-8: Variation in ST with variation in AC  
 a) Glottal source pulses  
 b) Corresponding magnitude frequency response

### 5.3.3 Controlling the Harmonic Richness Factor (HRF)

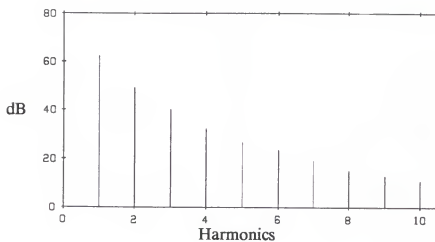
The Harmonic Richness Factor (HRF) is defined as the ratio of the sum of the magnitude of the harmonics to the magnitude of the fundamental frequency component in the glottal flow spectrum [Lee and Childers, 1989]. Normally, the magnitude of the fundamental frequency and the first nine harmonics are considered in the calculation of this glottal factor. The HRF is given by

$$HRF = \frac{\sum_{n=2}^{10} H_n}{H_1} \quad 5-1$$

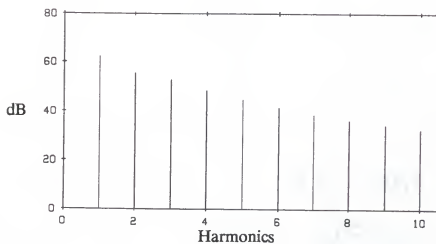
where  $H_n$  is the magnitude of the  $n$ th harmonic and  $H_1$  is the magnitude of the fundamental frequency component.

The Figure 5-9 shows the magnitude of the first ten harmonics in the magnitude frequency response of the glottal source pulses obtained from breathy and creaky phonations. It can be observed that the HRF is related to the shape of the magnitude frequency response of the glottal flow pulses in the low frequency region. The higher the value of the HRF, the flatter is the spectral slope in the low frequency region. The lower the value of the HRF, the steeper is the spectral slope in the low frequency region. It is assumed that the low frequency harmonics are not as much affected by additive noise and pitch perturbation as are the high frequency harmonics. Therefore, the HRF is considered to be determined by the voicing source pulses.

When the LF model is used as a voicing source model, the HRF depends upon the parameters  $t_p$ ,  $t_e$  and  $t_a$ , and thus, upon time domain glottal factors, such as OQ, SQ and AC. The variation in the value of the HRF due to the variation in each of these time domain glottal factors is graphically illustrated in Figure 5-10. From these graphs we can deduce the value of the HRF for a combination of the values of OQ, SQ and AC, or, vice versa. It can be observed from these figures that HRF is more sensitive to OQ and SQ than it is to AC. The insensitivity to AC is expected since AC

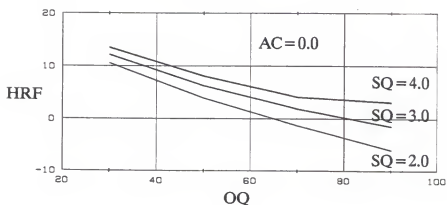


[a]

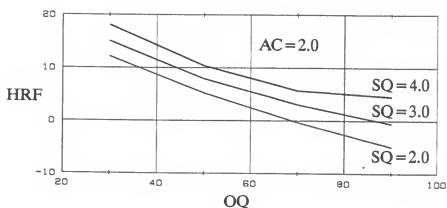


[b]

Figure 5-9: Magnitude of fundamental and first nine harmonics of glottal flow pulses for  
a) Breathly phonation (HRF=0.428)  
b) Creaky phonation (HRF=1.584)

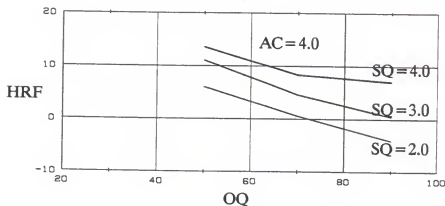


[a]



OQ

[b]



OQ

[c]

Figure 5-10: Variation in HRF due to variation in OQ, SQ and AC  
 a)  $AC = 0.0$   
 b)  $AC = 2.0$   
 c)  $AC = 4.0$  (OQ = 30, 40 not possible for SQ = 2.0, 3.0 and 4.0)



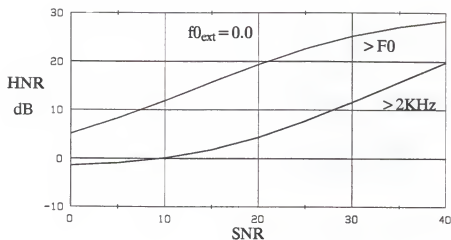
determines the spectral slope of the magnitude frequency response in the high frequency region and has little effect on the spectral slope in the low frequency region.

#### 5.3.4 Controlling the Harmonic to Noise Ratio (HNR)

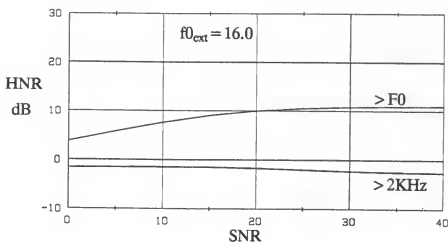
The Harmonic to Noise Ratio (HNR) is defined as the ratio of power in the harmonics to the power in the inter-harmonic components. This ratio measures the degradation of harmonics, which is commonly observed in the spectrum of the speech signal for breathy, rough and hoarse voices. The higher the degradation of the spectrum, the lower is the value of the HNR. In this study we have defined the HNR measure as the degradation of the spectrum of the glottal flow pulses obtained by inverse filtering speech signal. The definition and the characteristics of the HNR measure are discussed in Appendix E. In this chapter we discuss only the relationship of the HNR measure to the several parameters of the new glottal source model.

The HNR measure can be used as one of the significant glottal factors for modeling/synthesizing various vocal characteristics. From a modeling/synthesis point of view, it can be hypothesized that the degradation of harmonics in the speech spectrum is due to the presence of high-frequency aspiration noise and pitch period perturbation in the glottal flow pulses. When simulating glottal flow pulses by a glottal source model, it may be assumed that the voicing source pulses represent a “smooth” signal. The spectrum of the “smooth” voicing source pulses consists only of the fundamental frequency component and its harmonics. The additive aspiration noise and the pitch perturbation cause the degradation of harmonics and inter-harmonic components. By controlling the power in the aspiration noise, i.e., the Signal to Noise Ratio (SNR), and the extent of the pitch perturbation, “ $f0_{ext}$ ,” we can synthesize speech tokens with varying degrees of degradation in the glottal source spectrum. The speech tokens synthesized from such glottal source pulses should have a varying degree of breathiness, roughness and hoarseness.

The Signal to Noise Ratio (SNR) is determined by the ratio of the signal power to the noise power in the glottal source pulses. The SNR of the glottal source pulses can be controlled by the glottal source model's parameters "av" and "ah." In Figure 5-11 a variation of the HNR due to a variation in SNR (signal power constant and noise power variable) for different values of the extent of pitch perturbation, " $f0_{ext}$ ," is shown. In the Figure 5-12 a variation of the HNR due to variation in the extent of pitch perturbation, " $f0_{ext}$ ," for different values of SNR is shown. It can be observed that the value of the HNR decreases as the SNR decreases and as the extent of pitch perturbation increases. From the Figure 5-11b, it can be observed that when  $f0_{ext} \neq 0$  and SNR is varied from a low value to a high value, the HNR is initially determined by the value of " $f0_{ext}$ " when SNR is small, and then the HNR is increasingly affected by SNR as its value increases. A similar observation can be made from Figure 5-12b when  $SNR \neq \infty$  and  $f0_{ext}$  is varied from a low to a high value. The value of HNR is initially determined by the value of SNR when  $f0_{ext}$  is small, and then the HNR is increasingly determined by  $f0_{ext}$  as its value increases. The HNR is sensitive to the shape of the spectrum of the aspiration noise. The variation in HNR due to a variation in the filter coefficient, " $a_n$ " and the type of filter is shown in Figure 5-13a. The HNR is also sensitive to the zero crossing rate of change of pitch perturbation. The variation in HNR due to a variation in the "rate of pitch perturbation" obtained by variation in the filter coefficient, " $a_{\theta}$ " and the type of filter is shown in Figure 5-13b. It can be observed that the bandwidth of a first order highpass filter does not change appreciably when the filter coefficient is varied from -1.0 to 0, and hence, the HNR does not vary appreciably. A second or higher order filter that allows a wide range of variation in the bandwidth of the passband may be required for a better control of HNR via " $a_n$ " and " $a_{\theta}$ " parameters.

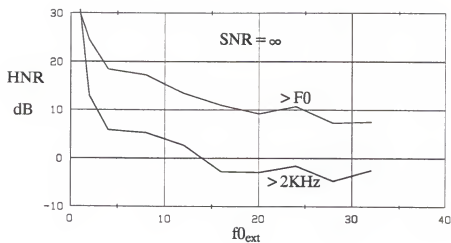


[a]

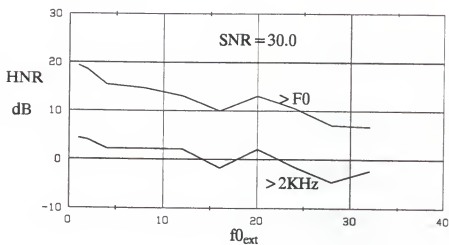


[b]

Figure 5-11: Variation in HNR with variation in SNR  
 $> F0$ : HNR measured from harmonics greater than  $F0$  and  
 $> 2KHz$ : HNR measured from harmonics greater than 2 KHz  
 a) extent of pitch perturbation  $f0_{ext} = 0.0$   
 b) extent of pitch perturbation  $f0_{ext} = 16.0$

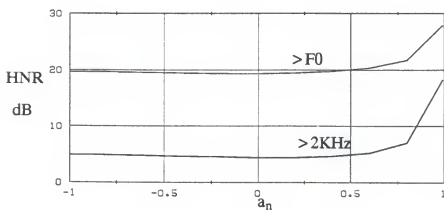


[a]

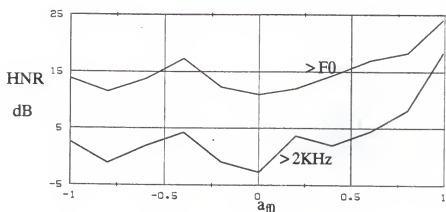


[b]

Figure 5-12: Variation in HNR with variation in  $f0_{ext}$   
 $> F0$ : HNR measured from harmonics greater than  $F0$  and  
 $> 2KHz$ : HNR measured from harmonics greater than 2 KHz  
a) extent of pitch perturbation  $SNR = \infty$   
b) extent of pitch perturbation  $SNR = 30.0$



[a]



[b]



Figure 5-13: Variation in HNR with variation in  
 $> F0$ : HNR measured from harmonics greater than  $F0$  and  
 $> 2KHz$ : HNR measured from harmonics greater than 2 KHz  
 a) spectrum of aspiration noise,  $a_n$  (SNR = 30.0)  
 b) rate of pitch perturbation,  $a_{f0}$  ( $f0_{ext} = 16.0$ )

#### 5.4 Frequency Domain Approach to Modeling Vocal Disorders

In this section we describe a frequency domain approach to modeling various vocal characteristics. Since the acoustical characteristics of the speech signal are perceptually closer to the frequency domain characteristics than the time domain characteristics of the speech signal, we have hypothesized that the frequency domain approach to modeling various vocal characteristics may be superior to the time domain approach. Fant and Lin, (1988) have recently used the frequency domain approach to model normal and breathy voices. In their method, the LF model's parameters are directly obtained from the spectra of the speech signal. In our approach we specify the new glottal source model's parameters to obtained glottal source pulses with desired frequency domain characteristics.

The following approaches are suggested for modeling various vocal characteristics in terms of frequency domain glottal factors.

- 1) Matching the spectra of glottal flow pulses with the magnitude frequency response of glottal source pulses generated by a glottal source model. Using this procedure we can analyze the frequency domain characteristics of glottal flow pulses that are significant for modeling various vocal characteristics.
- 2) Generating the glottal source pulses with the desired frequency domain characteristics from a glottal source model. By systematically varying the frequency domain glottal factors and by finding the frequency domain glottal factors significant of various voice types through listening tests, we can obtain models for various vocal characteristics. This approach has the merit that it does not have the problems associated with data collection and inverse filtering of speech signal.

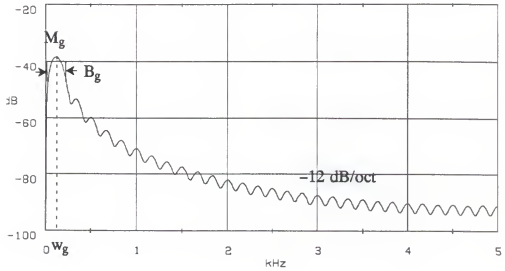
For coarse simulation of the frequency domain glottal factors in the magnitude frequency response of glottal source pulses, it is adequate to control the frequency domain glottal factors, such as  $F_0$ , ST, HRF and HNR. The glottal factor  $F_0$

determines the number of harmonics in the magnitude frequency response. The glottal factors HRF and ST jointly determine the amplitude of the low-frequency harmonics relative to the amplitude of the mid- and high-frequency harmonics. The ST determines the amplitude of high-frequency harmonics. The HNR determines the level of inter-harmonic components with respect to the harmonics in the magnitude frequency response. However, if a detailed simulation of the frequency domain glottal flow characteristics is required, i.e., specification of a detailed shape of the magnitude frequency response of glottal source pulses is required, we need to define additional frequency domain glottal factors.

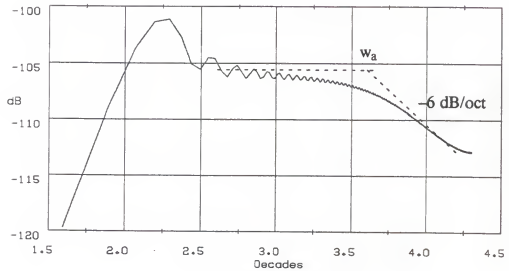
We have defined additional frequency domain glottal factors based upon a typical shape of the magnitude frequency response of a single differentiated glottal source pulse. The Figure 5-14 illustrates these additional frequency domain glottal factors. These glottal factors are defined as follows:

- 1)  $M_g$  : Amplitude of the peak of the magnitude frequency response of a differentiated glottal source pulse.
- 2)  $w_g$  : The frequency at which the peak of the magnitude frequency response of a differentiated glottal source pulse occurs.
- 3)  $B_g$  : Width (bandwidth) of the peak of the magnitude frequency response of a differentiated glottal source pulse.
- 4)  $w_a$  : Corner frequency at which the spectral tilt changes from  $-6$  dB/oct to  $-12$  dB/oct in the magnitude frequency response of a differentiated glottal source pulse.

It can be observed that the previously defined glottal factors and these four glottal factors together determine a detailed shape of the magnitude frequency response of a differentiated glottal source pulse. The  $M_g$ ,  $w_g$  and  $b_g$  glottal factors determine the shape of the magnitude frequency response in the low frequency region, specifically the shape of the main lobe, and the fourth glottal factor " $w_a$ " determines the level of the harmonics in the mid- and high-frequency region of the magnitude frequency



[a]



[b]

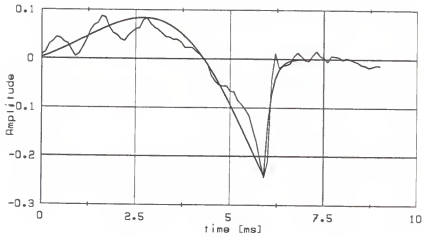
Figure 5-14: Frequency domain glottal factors of voicing source pulses  
 a) Magnitude frequency response of a differentiated glottal source pulse  
 b) Magnitude frequency response of a double differentiated glottal source pulse. (The X-axis is represented in terms of log of frequency. The value of 10 to the power "Decades" corresponds to the frequency in Hz.)



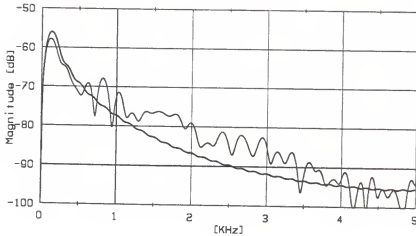
response. These glottal factors can be used to specify additional details of the shape of the magnitude frequency response.

When the LF model is used as a voicing source, the LF model's "timing parameters" can be used to control these frequency domain glottal factors. The relationships between the "timing parameters" of the LF model and these frequency domain glottal factors are described in a technical report "The LF model" [Lalwani, 1991]. The results of several experiments to find the trend in the variations in the values of these frequency domain glottal factors due to variations in the values of the "timing parameters" are given in this report. Also, the procedure for modeling various vocal characteristics via: 1) time and frequency domain matching of the inverse filtered differentiated glottal source pulses with the differentiated glottal source pulses generated by LF model, and 2) systematically varying the frequency domain glottal factors in the differentiated glottal source pulses generated by the LF model are also described in this technical report.

We tested our procedure for time and frequency domain matching of the differentiated glottal flow pulse obtained by inverse filtering of speech signal of a breathy phonation with the differentiated glottal source pulse generated by the LF model. In the Figure 5-15 the time domain matching of the differentiated glottal flow pulse obtained by inverse filtering the speech signal for a single pitch period is shown. The comparison of the magnitude frequency response show differences around the peak and in the mid-frequency region. In Figure 5-16 the frequency domain matching of the same differentiated glottal flow pulse is shown. Note that the magnitude frequency response shows better matching in the peak and mid-frequency region than observed in Figure 5-15b. However, the comparison of waveforms in Figure 5-16a show a mis-match at the zero-crossing in the middle of the differentiated glottal source pulse. It can be observed that the differentiated glottal flow pulse has a low frequency ripple, which causes a deviation from the actual values of the "timing

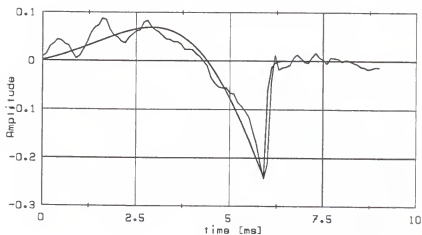


[a]

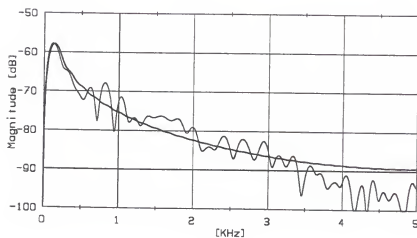


[b]

Figure 5-15: Time domain matching  
 a) Differentiated glottal flow pulse and glottal source pulse  
 b) Corresponding magnitude frequency response  
 (The LF model parameters are  $E_c = -0.23$ ,  $t_p = 4.4\text{ms}$ ,  
 $t_c = 6.0\text{ms}$ ,  $t_a = 0.17$ )



[a]

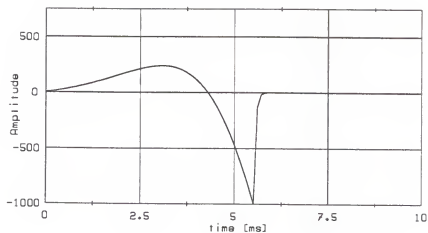


[b]

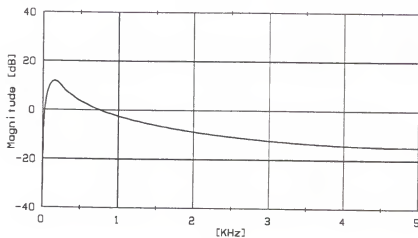
Figure 5-16: Frequency domain matching  
 a) Differentiated glottal flow pulse and glottal source pulse  
 b) Corresponding magnitude frequency response  
 (The LF model parameters are  $E_c = -0.23$ ,  $t_p = 4.6\text{ms}$ ,  
 $t_c = 6.0\text{ms}$ ,  $t_a = 0.08$ )

parameters,” such as the instants of zero-crossing, negative peak, closure, etc. By performing the frequency domain matching, specifically by matching the peak of the magnitude frequency response with the peak of the spectrum of differentiated glottal flow pulse, such low frequency trends can be easily de-emphasized and the “timing parameters” can be estimated more accurately. Another advantage is that the frequency domain matching gives better estimates of “timing parameters” than the time domain matching if localized artifacts are present in the glottal flow waveform at the significant timing instances. The significance of frequency domain matching is evaluated in a better way by listening to synthetic speech tokens. The speech tokens synthesized with the frequency domain matched glottal source pulse (in Figure 5-16a) as glottal source sounded more similar to the speech tokens synthesized with the original differentiated glottal flow pulse as glottal source than the speech tokens synthesized with the time domain matched glottal source pulse (in Figure 5-15a) as glottal source.

The frequency domain approach to modeling vocal characteristics involves systematic variation of frequency domain glottal factors in glottal source pulses. However, most of the frequency domain glottal factors cannot be directly controlled. Instead they must be controlled by systematic variations in the time domain glottal factors. The “initial” (“pilot”) values of the time domain glottal factors for such variations can be estimated from the tables listed in Gobl, (1989), Lee (1988) and/or from Table 4-I. Figure 5-17, Figure 5-18 and Figure 5-19 show a differentiated voicing source pulse and its magnitude frequency response obtained from the “pilot” values of the time domain glottal factors for creaky, modal and breathy voices, respectively. The “pilot” values for the variation of the frequency domain glottal factors can be estimated from the magnitude frequency response for each vocal characteristics.

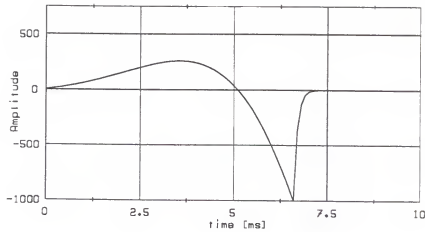


[a]

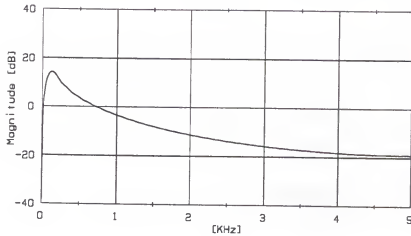


[b]

Figure 5-17: Creaky vocal characteristic  
 a) Differentiated voicing source pulse  
 b) Corresponding magnitude frequency response  
 (The LF model parameters are  $E_c = 60\text{dB}$ ,  $t_p = 4.4\text{ms}$ ,  
 $t_e = 5.6\text{ms}$ ,  $t_a = 0.1$ ,  $t_0 = 23.3\text{ms}$ )

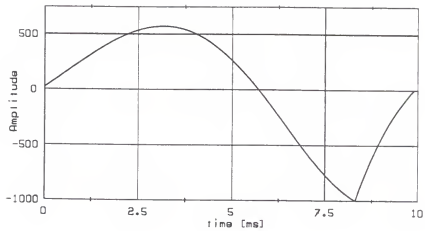


[a]

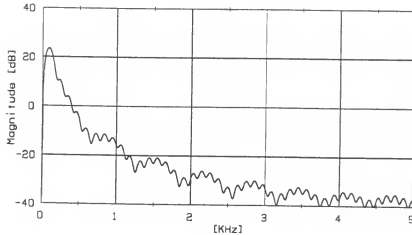


[b]

Figure 5-18: Modal vocal characteristic  
 a) Differentiated voicing source pulse  
 b) Corresponding magnitude frequency response  
 (The LF model parameters are  $E_c = 60\text{dB}$ ,  $t_p = 5.2\text{ms}$ ,  
 $t_e = 6.7\text{ms}$ ,  $t_a = 0.1$ ,  $t_0 = 11.1\text{ms}$ )



[a]



[b]

Figure 5-19: Breathy vocal characteristic  
 a) Differentiated voicing source pulse  
 b) Corresponding magnitude frequency response  
 (The LF model parameters are  $E_c = 60\text{dB}$ ,  $t_p = 5.8\text{ms}$ ,  
 $t_e = 8.4\text{ms}$ ,  $t_a = 10.0$ ,  $t_0 = 10.0\text{ms}$ )

### 5.5 Summary

The time domain glottal factors that are significant for modeling various vocal characteristics can be precisely and independently controlled by using the new glottal source model. Also, these glottal factors can be directly specified through the new glottal source model's parameters. The frequency domain glottal factors are related to the time domain glottal factors. Not all the frequency domain glottal factors can be directly specified through the new glottal source model's parameters or can be controlled independent of one another. For such frequency domain glottal factors we have illustrated their relationship with the related time domain glottal factors graphically.

We have proposed a frequency domain approach to modeling various vocal characteristics. For coarse simulation of the frequency domain glottal flow characteristics in the glottal source pulses, it is adequate to control the frequency domain glottal factors, such as  $F_0$ , ST, HRF and HNR in the magnitude frequency response. However, for a detailed simulation of frequency domain characteristics, precise control of frequency domain glottal factors is required. We have defined additional frequency domain glottal factors, such as  $M_g$ ,  $w_g$ ,  $B_g$  and  $w_a$  in addition to the glottal factors  $F_0$ , ST, HRF and HNR that can be used to specify a detailed shape of the magnitude frequency response of the glottal source pulses. We have also developed the procedures to manipulate the values of the frequency domain glottal factors in order to 1) match the magnitude frequency response of the differentiated glottal flow pulses obtained by inverse filtering of the speech signal with that of the differentiated glottal source pulses generated by a glottal source model, and 2) systematically vary the frequency domain characteristics of the glottal source pulses generated by the LF model. Using these procedures it is possible to obtain models for various vocal characteristics in terms of frequency domain glottal factors.



## CHAPTER 6 EXPERIMENTS AND PERCEPTUAL EVALUATIONS

### 6.1 Introduction

In this chapter we describe the experiments to test the performance of the new glottal source model. These experiments involved a systematic variation of the time domain glottal factors. These experiments were designed to validate certain hypotheses concerning the relationships between the time domain glottal factors and various vocal characteristics. Similar experiments may be designed to validate other hypotheses.

### 6.2 Experiments

The new glottal source model has been incorporated in the flexible formant synthesizer. In a formant synthesizer the glottal source pulses are generated by specifying the glottal source parameters. The excitation to the filter banks is provided by the glottal source pulses to obtain synthetic speech. By specifying appropriate values for the new glottal source model's parameters, it is possible to generate glottal source pulses with glottal flow characteristics typical of a vocal characteristic. The synthetic speech tokens generated from these glottal source pulses have perceptual qualities similar to that particular vocal characteristic. We conducted several experiments with the new glottal source model and the flexible formant synthesizer to synthesize speech tokens with various vocal characteristics. These experiments are described in the following sub-sections.

### 6.2.1 Objectives

- 1) To evaluate the performance of the new glottal source model for synthesizing various vocal characteristics.
- 2) To determine the significance of each of the time domain glottal factors in modeling various vocal characteristics.
- 3) To determine the effect of variation of one or more of the time-domain glottal factors on the perception of each vocal characteristics.

### 6.2.2 Method

For each vocal characteristics:

I) The value of each glottal factor was estimated from the Table 4-I. The values of the appropriate glottal source model's parameters were estimated from the estimates of glottal factors.

II) Several sets of sequences of glottal source pulses with progressively varying glottal factors were generated. In each set, the value of the glottal factors under investigation was systematically varied across sequences but kept constant for each sequence. The following procedure was used to generate each set of sequences:

- 1) A set of glottal source pulses was generated for each of the voicing source related glottal factors, such as fundamental frequency, open quotient, speed quotient and abruptness of closure. The estimated values of the glottal factor under investigation was treated as "pilot" value for variation. For the "rough" and "hoarse" vocal characteristics the "preferred" values of the voicing source related glottal factors were hypothesized to be the same as those obtained for "modal" voice. Go to step 6.

- 2) A set of glottal source pulses was generated for each of the aspiration noise source related features such as the power, spectrum and amplitude-modulation. The voicing source related glottal factors were kept constant at their respective "preferred" values while generating this set of sequences of glottal source pulses. Go to step 6.
- 3) A set of glottal source pulses was generated for each of the pitch perturbation features such as "the extent of pitch perturbation" and "the rate of pitch perturbation." As mentioned earlier, the voicing source related glottal factors, such as open quotient, speed quotient and abruptness of closure remain constant at their respective "preferred" values even if the fundamental frequency of the voicing source pulses is varied by the pitch perturbation source. Go to step 6.
- 4) A set of glottal source pulses was generated for each of the aspiration noise source features, such as the power, spectrum and amplitude-modulation, in the manner similar to variations carried out in step 2. The voicing source related glottal factors and the pitch perturbation source related parameters were kept constant at their respective "preferred" values. Go to step 6.
- 5) A set of glottal source pulses was generated for each of the pitch perturbation source related parameters in the manner similar to variations carried out in step 3. The voicing source related glottal factors and the aspiration noise source related features such as the power, spectrum and amplitude-modulation were kept constant at their respective "preferred" values. Go to step 6.
- 6) From each set of sequences of voicing/glottal source pulses, a set synthetic speech tokens of the sustained vowel /I / was generated using the cascade filter bank in the flexible formant synthesizer. Each speech token had a duration of 2 seconds. The formant structure for vowel /I / was not varied

during the synthesis. The overall intensity of speech tokens were adjusted to the same level. (However, the perceived loudness were not necessarily the same since different voice samples may have different energy distributions in frequency.)

7) Separate listening test were carried out for each glottal factor under investigation. The judges for the listening test were three professors, two from Speech Department and one from the Electrical Engineering Department who was an experienced speech scientist. Each set of speech tokens generated in step 6 was presented to the trained speech researchers via headphones in a random order in a presentation sequence. The speech tokens in each presentation sequence were randomly arranged. The presentation sequence could be repeated if any of the judges wished to listen the speech tokens again. The judges were asked to analyze the "quality," i.e., naturalness, of the vocal characteristics of the speech token. They were asked to indicate the speech tokens whose "quality" they preferred and suggest the variation(s) in the glottal factor(s) in order to improve the "quality" of the other speech tokens. The values of the glottal factors for the "preferred" speech tokens were stored as "preferred" values for the vocal characteristic being synthesized. Continue this procedure up to step 5.

III) The sentence "We were away a year ago" was synthesized from the glottal source pulses generated with the "preferred" values of the glottal factors for each vocal characteristics. During the synthesis of the sentence, the glottal source parameters were kept constant except for the fundamental frequency and the voicing gain parameters. The same formant tracks were used while synthesizing the sentence for each vocal characteristic.

### 6.2.3 Results

The listeners' evaluation of the "quality," i.e., naturalness, of the vocal characteristics of the synthesized speech tokens are summarized in this sub-section. The speech tokens synthesized from the sequence of voicing source pulses with the "preferred" values of the voicing source related glottal factors for each of the modal, creaky and breathy voices had the perceptual characteristics of these voices, but lacked naturalness and sounded monotonic. There was improvement in the naturalness and reduction of monotony when the speech tokens were synthesized from the glottal source pulses generated by adding aspiration noise to the voicing source pulses with perturbation of the fundamental frequency. The perturbation of fundamental frequency was more significant for the reduction of monotony of the synthesized speech than was the addition of aspiration noise to the voicing source pulses. The speech tokens synthesized from the glottal source pulses, for which the values of voicing source related glottal factors were close to the "preferred" values for the "modal" vocal characteristics, and the aspiration noise and perturbation of fundamental frequency was higher than the "preferred" values for the "modal" vocal characteristics, had the perceptual characteristics of "rough" and "hoarse" voices. In general, the listeners preferred speech tokens synthesized from glottal source pulses generated by adding aspiration noise to the vocal source pulses with perturbation of the fundamental frequency. The "preferred" values of the glottal factors and the new glottal source model's parameters for each vocal characteristic are listed in Table 6-I.

The listeners' evaluations related the glottal factors to various vocal characteristics. The speech tokens synthesized from the glottal source pulses had the following characteristics:

- 1) low fundamental frequency sounded creaky,
- 2) high open quotient gave a perception of "lax" voice,

Table 6-1  
Preferred values of the glottal factors for various vocal characteristics

| Parameter | $t_0$<br>ms | OQ<br>% <sup>1</sup> | SQ<br>% <sup>1</sup> | $t_a$<br>% <sup>1</sup> | SNR<br>(dB) | JF<br>% <sup>2</sup> |
|-----------|-------------|----------------------|----------------------|-------------------------|-------------|----------------------|
| Quality   |             |                      |                      |                         |             |                      |
| Modal     | 10          | 61.5                 | 3.0                  | 0.5                     | 70.0        | 0.82                 |
| Creaky    | 25          | 20.3                 | 3.0                  | 0.1                     | 80.0        | 2.46                 |
| Breathy   | 10          | 100.0                | 2.0                  | 8.0                     | 55.0        | 0.82                 |
| Rough     | 12.5        | 61.5                 | 3.0                  | 0.5                     | 60.0        | 1.62                 |
| Hoarse    | 5.71        | 61.5                 | 3.0                  | 0.5                     | 45.0        | 1.62                 |

Preferred values of the new glottal source model parameters for various vocal characteristics

| Parameter | f0<br>Hz | $t_p$<br>% <sup>1</sup> | $t_c$<br>% <sup>1</sup> | $t_a$<br>% <sup>1</sup> | av<br>(dB) | ah<br>(dB) | amp1 | amp2 | offset<br>% <sup>1</sup> | dur<br>% <sup>1</sup> | a <sub>n</sub> | a <sub>0</sub> | f0 <sub>ext</sub><br>% <sup>2</sup> | f0 <sub>tp</sub><br>% <sup>2</sup> |
|-----------|----------|-------------------------|-------------------------|-------------------------|------------|------------|------|------|--------------------------|-----------------------|----------------|----------------|-------------------------------------|------------------------------------|
| Quality   |          |                         |                         |                         |            |            |      |      |                          |                       |                |                |                                     |                                    |
| Modal     | 100      | 45.0                    | 60.0                    | 0.5                     | 100        | 63.0       | 0.0  | 1.0  | 0.5                      | 0.5                   | -1.0           | 0.0            | 5.0                                 | 15.0                               |
| Creaky    | 40       | 15.0                    | 20.0                    | 0.1                     | 100        | 53.0       | 0.0  | 1.0  | 0.5                      | 0.5                   | -1.0           | 0.0            | 15.0                                | 15.0                               |
| Breathy   | 100      | 60.0                    | 90.0                    | 8.0                     | 100        | 78.0       | 0.0  | 1.0  | 0.5                      | 0.5                   | -1.0           | 0.0            | 5.0                                 | 15.0                               |
| Rough     | 80       | 45.0                    | 60.0                    | 0.5                     | 100        | 73.0       | 0.0  | 1.0  | 0.5                      | 0.5                   | -1.0           | 0.0            | 10.0                                | 15.0                               |
| Hoarse    | 175      | 45.0                    | 60.0                    | 0.5                     | 100        | 88.0       | 0.0  | 1.0  | 0.5                      | 0.5                   | -1.0           | 0.0            | 10.0                                | 15.0                               |

%<sup>1</sup> : percent of pitch-period

%<sup>2</sup> : percent of fundamental frequency

- 3) low speed quotient gave a perception of "filtered" voice.
- 4) high speed quotient gave a perception of "tense" voice,
- 5) aspiration noise gave a perception of breathiness,
- 6) high frequency aspiration noise gave a perception of breathiness,
- 7) amplitude-modulated aspiration noise source were more "preferable,"
- 8) low or medium values of pitch perturbation were "preferable;" high pitch perturbation resulted in "noisy" speech tokens,
- 9) lowpass filtered pitch perturbation sounded "unnatural" and,
- 10) highpass filtered pitch perturbation sounded "rough."

#### 6.2.4 Conclusions

The above experiments indicate that our glottal source model has the potential for synthesis of speech with various vocal characteristics. In general, we hypothesized that the glottal source model for each of these vocal characteristics should include both an aspiration noise source and a pitch perturbation source. The degree to which each of these two sources is used may depend upon the vocal characteristic to be synthesized. Further experiments, which may also involve systematic variation of the frequency domain glottal factors, should be conducted for improving the glottal source model and to establish typical values of both the time and frequency domain glottal factors required for modeling each vocal characteristic.

## CHAPTER 7 DISCUSSION

### 7.1 Summary

This research investigated several aspects of formant speech synthesis. First, a flexible formant synthesizer was developed. Flexibility was incorporated in the synthesizer parameter specification, synthesis algorithm and synthesizer architecture. A new glottal source model was developed and incorporated in the flexible formant synthesizer for improving the quality of synthetic speech. Various glottal factors significant for synthesizing/modeling different voice types could be controlled through this glottal source model's parameters. With appropriate combinations of these glottal factors, i.e., the glottal source characteristics, various voice types could be synthesized. This study demonstrated the feasibility of modeling various voice types through synthesis.

#### 7.1.1 The Flexible Formant Synthesizer

The flexible formant synthesizer is an outgrowth of Klatt's cascade/parallel formant synthesizer. We have enhanced Klatt's synthesizer by improving parameter specification procedures, incorporating many new parameters and modifying the synthesis algorithms and the synthesizer architecture. These enhancements have resulted in an increase in the flexibility and efficiency of the synthesizer.

We have improved the specification of: 1) glottal source model and noise source model, 2) first order filters in the synthesizer architecture for spectral-shaping of the



glottal and noise excitation source, 3) configuration of filter banks, 4) duration of synthesis of a speech utterance, 5) type of synthesis, such as pitch-synchronous or fixed-frame synthesis, 6) simulation of source-tract interaction, etc.

We have incorporated several parametric and non-parametric glottal source models in the flexible formant synthesizer. These glottal source models are: 1) Klatt's model, 2) three/two pole model, 3) LF model, 4) new glottal source model, 5) impulse train, 6) single pulse waveform and 7) multi-pulse waveform. Provisions have been made to add other glottal source models. The user can select any one glottal source model as a voicing source for synthesis of voiced speech sounds. A built-in random number generator or external random number tables can be used as excitation sources for synthesis of unvoiced sounds. A conjunction of these two sources is used for synthesis of mixed sounds.

We have developed FOS (First Order System) for adding flexibility in the synthesizer architecture. These FOS systems can be used either as a first order lowpass or highpass filter to modify the spectra of glottal source, noise source and output signal. The value of the filter coefficient of a FOS, in addition to determining the bandwidth of the passband of a filter, also determines the type of first order filter it simulates. A positive value simulates a lowpass first order filter whereas a negative value simulates a highpass first order filter.

We have incorporated flexibility in specifying filter parameters, synthesis algorithm and synthesizer architecture in order to create a flexible configuration of the cascade and the parallel filter banks. The dynamic configuration of the filter banks at the start-up and also at each speech frame boundary enables us to have a variable number of filters in the cascade and parallel filter banks at start-up and during synthesis. With the flexible formant synthesizer, we can synthesize speech from the exact number of continuous and discontinuous formants and anti-formants specified for synthesis (without inserting default values for the formants and anti-formants that

are not specified). The algorithm for dynamic configuration of filter banks also includes detection and removal of high-amplitude short duration transients from synthetic speech. The synthesized speech is free of "clicks" and "pops" even if the formant tracks are unsmoothed and/or discontinuous.

The flexible formant synthesizer can be used in three different possible configurations: 1) all-cascade, 2) all-parallel and 3) cascade/parallel synthesizer configuration. In order to use the synthesizer in the all-parallel configuration, the parallel filter bank should simulate the magnitude frequency response of the cascade filter bank during the synthesis of voiced sounds. We have improved Klatt's procedure for simulating the cascade filter bank by a parallel filter bank and also developed an entirely new procedure for the same. Also, by appropriate filter bank specification we can simulate "zeros" in the magnitude frequency response of the parallel filter bank. Our method can be used even with the flexible configuration of the cascade and parallel filter banks.

We have incorporated appropriate parameters and algorithms for time and frequency scaling of the synthetic speech signal. The time scaling is achieved by skipping/repeating the portions of parameter tracks during the synthesis; we do not have to separately modify the signal parameter tracks. The frequency scaling is achieved by scaling the formant frequencies and bandwidths and the sampling rate of the synthesis. The flexible formant synthesizer synthesizes a smooth speech signal even if the formant tracks are abruptly changed during the time and frequency scaling of the synthesized speech signal.

In the flexible formant synthesizer, we have made provision for simulating source-tract interaction. A glottal source model, such as the new glottal source model, can be used to generate glottal source pulses with right-skewness. The first formant bandwidth and frequency can be changed during the open phase of the glottal source pulse to simulate the effect of truncation of the first formant oscillations during the

open portion of the glottal source pulses. We observed that changing the first formant bandwidth incrementally provides a better simulation of the first formant truncation effect than other methods. We have described strategies for synthesizing sustained vowels and sentences with voiced, unvoiced and mixed sounds. The new/modified features of the flexible formant synthesizer are illustrated with several examples of synthesis of sustained vowels and sentences.

### 7.1.2 New Glottal Source Model

Our approach to modeling various voice types is via formant speech synthesis. By controlling the source and vocal tract characteristics, independently of one another, we can synthesize speech tokens with various vocal characteristics. Our hypothesis is that various voice types are the symptoms of laryngeal dysfunctions and can be synthesized by controlling various characteristics of glottal source pulses (glottal factors).

In this study we focused on the vocal characteristics of modal, creaky, breathy, rough and hoarse voice types. In the literature we observed that the time domain glottal factors, such as pitch-period, glottal pulse width, glottal pulse skewness, abruptness of closure of the glottal pulse, aspiration noise, jitter and shimmer, and the frequency domain glottal factors, such as spectral tilt, Harmonic Richness Factor and Harmonic to Noise Ratio were described as significant for modeling these voice types. We developed a new glottal source model such that these glottal factors can be controlled in the glottal source pulses through the glottal source model's parameters. This model is comprised of 1) the voicing source (LF model), 2) the aspiration noise source, 3) the pitch perturbation source and 4) the amplitude perturbation source. The voicing source generates voicing source pulses that simulate the "smoothed" glottal flow pulses generated by quasi-periodic vibration of vocal folds. The aspiration noise source simulates the additive noise component of the

glottal flow pulses. With appropriate amplitude-modulation, it can also be used to simulate turbulent air flow generated at the glottis due to incomplete closure. The pitch perturbation source causes perturbation of the fundamental frequency (pitch-period) parameter to simulate aperiodicity in the vocal fold vibration. The amplitude perturbation source simulates the variation in the peak amplitude of the glottal flow pulses by causing the perturbation of the voicing gain parameter.

The time domain glottal factors can be directly specified through the new glottal source model's parameters. Also, each time domain glottal factor can be controlled independently of other time domain glottal factors. The shape of the voicing source pulses determine the overall shape of glottal source pulses, and hence, the time domain glottal factors such as pitch period, glottal pulse width, glottal pulse skewness, abruptness of closure, are determined by the parameters of the voicing source model. The glottal factor SNR is determined by the rate of the power in the glottal source pulses to the aspiration noise. The extent of pitch perturbation in the voicing source pulses controls the "jitter" and the extent of amplitude perturbation in the voicing source pulses controls the "shimmer" in synthetic speech. We can synthesize speech tokens having the specified values of the pitch perturbation measures, such as the "Jitter Factor," "Frequency Perturbation Quotient" and "Directional Jitter."

The frequency domain parameters are related to the time domain parameters and cannot be directly specified through the parameters of the new glottal source model. The relationships between the time domain glottal factors and the frequency domain glottal factors are illustrated by analytical expressions and by graphs. The frequency domain glottal factors, such as fundamental frequency, spectral tilt and Harmonic Richness Factor are related to the shape of the glottal source pulses and hence can be controlled by the parameters of the voicing source model. The Harmonic to Noise Ratio is determined by the combination of the power in the aspiration noise source, extent of pitch-perturbation and the extent of amplitude perturbation.

We synthesized several speech tokens of a sustained vowel /t/, each with a duration of 2 seconds. Each of the time domain glottal factors was systematically varied while keeping the other time domain glottal factors constant for the glottal source pulses that generated these speech tokens. The results from the informal listening tests indicate that with an appropriate combination of glottal factors, or the glottal source characteristics, various voice types could be synthesized. By varying the relative contribution from each source we can synthesize various voice types with varying severity. The sentence "We were away a year ago" synthesized with the appropriate combination of glottal factors for a voice type had the perceptual characteristics of that voice type.

The frequency domain glottal factors, such as fundamental frequency, spectral tilt and Harmonic richness Factor are sufficient for a coarse simulation of frequency domain characteristics of various voice types. We have defined additional frequency domain glottal factors based upon the magnitude frequency response of the glottal source pulses, such as  $M_g$ ,  $w_g$ ,  $B_g$  and  $w_a$ , for a detailed simulation of frequency domain characteristics of various voice types in synthetic speech. The motivation for this approach stems from the fact that the perceptual characteristics of the acoustic signal are more closely related to its frequency domain characteristics than its time domain characteristics. Hence, the frequency domain modeling approach is superior to the time domain modeling approach. With this approach, various voice types can be quantitatively described in terms of the frequency domain glottal factors. Further research on this approach is proposed.

## 7.2 Future Work

The quality of synthetic speech is important for many commercial applications and speech research areas. The formant synthesizer can synthesize high-quality

speech and can be used for modeling various voice types. The results derived in this study have been encouraging, but there is much still to do. Further research is suggested in several aspects of formant speech synthesis and modeling various voice types.

### 7.2.1 Improvements in Flexible Formant Synthesizer

- 1) The noise source model used for synthesizing frication and aspiration is primitive. Also, the spectra of fricatives and plosives are poorly understood and are still poorly simulated. Therefore, the quality of fricatives and plosives is inferior. Better source models and modeling of the spectra of these sounds are required for high-quality synthesis of these sounds.
- 2) The parallel filter bank is used for synthesizing the fricatives, plosives and nasal sounds. Currently, the parallel filter bank cannot simulate the anti-resonances that are commonly observed in the spectra of these sounds in natural speech. We have attempted to create "anti-formants" in the magnitude frequency response of the parallel filter bank, but the "anti-formant" frequencies and bandwidths are dependent upon the frequencies and bandwidths of the adjacent formants. Better simulation of "anti-formants" in the spectra of these sounds is required for "high-quality" synthesis.
- 3) The source-filter model of speech production considers the glottal source and the vocal tract independent of one another. Although, the synthesizers based upon the source-filter model can synthesize high-quality speech, it has been shown by several researchers that further improvement in the quality of synthetic speech can be brought about by incorporating source-tract interaction in the synthesis process. In the flexible formant synthesizer, we simulate source-tract interaction by using glottal source models that generate right-skewed glottal source pulses and the first formant truncation by increasing the first formant bandwidth. However, the shape of the

glottal source pulses does not vary according to the variations in the formant frequencies (load); the variation in the source characteristics does not affect the magnitude of the first formant truncation. Also, the resonance characteristics of the sub-glottal system are neglected. All of these factors should be considered for improving the quality of synthetic speech.

4) The quality of synthetic speech depends upon the rules and strategies applied for synthesizing various sounds in connected speech. Improvement in the quality of synthetic speech can be brought about by improving these rules and strategies.

### 7.2.2 Improvements in the New Glottal Source Model

1) In this study we focused on the modeling of various voice types in terms of the time domain glottal factors. This study should be further expanded to model various voice types in terms of frequency domain glottal factors. In this study we have outlined the procedure for this approach.

2) More experiments with systematic variation(s) of individual and combined glottal factors should be carried out and formal listening tests should be conducted for obtaining the typical values of the glottal factors and the range of values the glottal factors necessary for modeling various voice types.

3) In connected speech, various prosodic patterns are used to express different types of statements. It may be hypothesized that various stress and intonation patterns may be correlated to some of the glottal source characteristics other than the fundamental frequency contour and timing, e.g., glottal pulse width, glottal pulse skewness, etc. Also, the glottal source characteristics may depend upon the speech sound to be produced. For example, the glottal source characteristics for vowel /ɪ/ may be different from those for vowel /a/. Thus, an important continuation of the research reported here is to study voice source dynamics for connected speech.

- 4) For any voice disorder, a pathological condition alters the physiology of speech production, which may results in an alteration of the acoustical characteristics of the speech signal. In this study we have focused on the differences in the acoustical characteristics for various voice types. This study should be expanded to derive the cause and effect relationships between the physiological features of voice production and various voice types as reflected in aspects of the acoustic signal. The cause and effect relationships between the acoustical characteristics and various voice types can be used to aid the development of the physiological models of various voice types.
- 5) The knowledge of glottal source characteristics is useful for speech synthesis. It is believed that the knowledge of glottal source characteristics would also benefit the applications of speech recognition and speaker verification.



## APPENDIX A FORMANT SYNTHESIZER ARCHITECTURE

### Literature Survey of Formant Synthesizers

Early developments in the "terminal-analog" synthesizer were made with analogue electrical networks by Stewart, Dudley, Riesz, etc. as early as 1939. The operation of these networks were manually controlled by trained researchers.

Lawrence (1953) implemented a formant synthesizer with a parallel filter bank of three second order resonators. The synthesizer was defined as an electrical network excited by a source. Two different types of sources were used for excitation: a series of impulses specified by amplitude and frequency for "larynx excitation" and a white-noise generator for "fricative excitation." For each resonator the amplitude, the damping factor (bandwidth) and the resonance frequency were specified. The variation in the phase response of the resonators was not considered significant since the auditory system could not distinguish between phase differences of the sustained sinusoidal components. The synthesized speech was intelligible and adequate for commercial telephony. It was observed that the relative amplitudes of different resonances did not affect the intelligibility of speech sounds. Also, it was observed that the variations in damping were unimportant, and the third and higher formants did not contribute much to intelligibility.

Fant (1956) defined the terms "formant frequency," "formant bandwidth" and "formant amplitude." The speech production system was defined in terms of excitation source, vocal-tract transfer function and lip radiation load. This synthesizer was modelled as a source-filter system that separated (de-coupled) the glottal

excitation (source) from the vocal tract (filter) and assumed no source-tract interaction. The resulting system was described as a linear time varying source-filter model of speech production. This model had a "terminal-analog" approach where the synthesizer was treated as a "black-box" with no relationship to physiological or acoustical characteristics of the human speech production system. The synthesizer was implemented as a cascade of resonators. It was observed that for vowels, the specification of formant bandwidths was correlated with a particular pattern of formant frequencies. The formant amplitudes could be calculated once formant frequencies, bandwidths and slope of the source spectrum were specified. Flanagan (1957) suggested that for vowels, Fant's configuration of a cascade of resonators was theoretically more correct than a parallel connection of resonators. An anti-formant ("zero" or a notch) was produced in between two formants ("poles" or peaks) when two resonators were connected in parallel. The vowel quality was not as natural, if the relative amplitudes were not correctly specified. The author argued that the spectra of consonant sounds displayed both formants and anti-formants and therefore the cascade connection of resonators, which could not produce anti-formants in the spectrum, was incapable of producing such sounds. The author suggested that a combination of cascade and parallel filter banks might be used to achieve an efficient "terminal-analog" system for both vowels and consonants.

Flanagan, et al. (1956) did a computer simulation of a formant synthesizer for a text-to-speech system. The synthesizer was implemented as a sampled-data system. It had three digital resonators in cascade. This cascade filter bank also included a resonator and anti-resonator ("pole-zero") pair for synthesizing nasal sounds. The energy storage elements in the digital resonators were simulated by ideal delays. The impulse invariant transform was used to convert analog frequency-domain specifications to digital frequency-domain specifications. The synthesizer parameters were automatically obtained from the spectrograms.

Rabiner (1968) developed a sampled data system for a formant synthesizer. The advantage of sampled data systems was the inherent higher-pole correction. The author argued that no special higher-pole correction network was needed as in the case of analogue implementations. The voice bar was produced by by-passing the formant filter system and adjusting the gain of the excitation source. The fricatives were produced by using a white-noise (random number generator) to excite a frequency-shaping network, which was a combination of a resonator pair, and anti-resonator (all with fixed center frequencies and bandwidths) and a resonator and anti-resonator pair with variable center frequencies and bandwidths. Another frequency-shaping network at the output of the filter banks consisted of a resonator and anti-resonator pair to provide a high and low frequency emphasis for synthesized speech and to account for the radiation characteristics. The excitation network for voiced fricatives consisted of a glottal source model and a resonator with the same specifications as the first formant resonator to simulate the effect of the most significant resonance of the cavity between the glottis and the constriction in the vocal tract on the noise source. The output of this resonator was rectified and used to modulate the noise source. The author reported that the phonemes /z/ (as in zoo) and /ʒ/ (as in azure) were identified correctly 100% of the time in formal listening tests. The author also discussed the merits and demerits of cascade and parallel realizations of the synthesizer. The author concluded that, for synthesis-by-rule, serial implementations were better, since the number of parameters required was less (formant amplitude parameters were not required). The advantages of parallel synthesizers were that the noise generated in a fixed point architecture propagates additively rather than multiplicatively. The second advantage was that the sonorants could be synthesized through independent control of the formant amplitudes. The disadvantage of parallel synthesizers was that the notches between the formants in the

spectrum of vowels produced by the parallel realization were perceptible and were a corrupting factor during the listening tests.

Gold and Rabiner (1968) discussed the digital (hardware) implementation of the synthesizer that was similar to an earlier computer implementation by Rabiner (1968). The effects of finite register length with different SNR for vowels were demonstrated both theoretically and experimentally by computer simulations. Rabiner, et al., (1971) described a hardware realization of a digital formant speech synthesizer. This synthesizer shared one arithmetic logic unit for all resonators and used 24 bits to process the digital signals internally. It produced speech in "real-time" at a sampling rate of 12.8 KHz.

Hoimes (1973) described a software implementation of a formant synthesizer. He used a parallel filter bank for synthesizing both vowels and consonants. A "nasal-pole" was simulated by a resonator in the "normal first formant region." The resonators (formant generators for the second, third and fourth formants were connected in parallel and their outputs were added with alternate polarities. The nasal formant and the first three formants were dynamically controlled and the fourth formants was held fixed. Lowpass filters were used to smooth the amplitude, bandwidth and frequency control parameters. For the normal synthesis process, the bandwidths were preset to constant values and increased for the synthesis of nasals and consonants. The fourth formant, normally in the high frequency region from 3600 Hz to 4000 Hz, was simulated by a broadband filter. A formant-shaping network was connected in series with each resonator in the parallel filter bank in order to prevent the variations in the formant frequency, bandwidth or amplitude of one resonator from affecting the amplitudes of the other formants. The excitation source for voiced sounds was the second derivative of the glottal source pulses and had a flat spectrum. This was designed for the dynamic range of the data passing through the resonators, since the excitation source for consonants was a noise source with flat spectrum. The final

-6dB/oct slope for the speech spectrum was achieved by an output shaping filter (a lowpass filter) with a bandwidth of 640 Hz. The spectral-tilts of -6 dB/oct in the low frequency region for voicing sounds was obtained by inserting a lowpass filter in series with the glottal source mode. This filter had a passband with -6dB/oct slope for the frequency range up to 640 Hz and an approximately flat stopband above 640 Hz. The excitation source for synthesizing mixed sounds consisted of a glottal source in conjunction with a noise source. The proportion of the glottal source and the noise source components in the mixed excitation for a resonator in the parallel filter bank was based upon the center frequency of the formant it was generating. The resonators with low formant frequency had a predominantly strong glottal source component while the resonators with high formant frequency had a predominantly strong noise component. The parameter tracks were specified in files which were read at execution time. When the inverse filtered glottal volume-velocity waveform was used as a glottal source, the resultant speech was perceived as "natural sounding."

Klatt (1980) described a software implementation of a formant synthesizer that was a combination of both the cascade and the parallel filter banks. The formant synthesizer could be configured as a cascade/parallel synthesizer or as an all-parallel synthesizer. The cascade filter bank had five resonators in series. Nasal sounds were simulated by an additional resonator and anti-resonator pair in the cascade filter bank and an additional resonator in the parallel filter bank. The parallel filter bank had a total of six resonators and a single multiplier (by-pass path). In the cascade/parallel synthesizer configuration the cascade filter bank was used for synthesizing voiced and aspirated sounds, and the parallel filter bank was used for synthesizing the fricatives. In the all-parallel synthesizer configuration, the first formant generator (resonator) had only the glottal source as an excitation source. The second, third and fourth formant generators had a mixture of the differentiated glottal source and the noise source as an excitation source. The differentiated glottal source was used so that the

magnitude frequency response of the parallel filter bank at the low frequencies was not affected by the low frequency portion of the magnitude frequency response of the resonators with high formant frequencies (Holmes, 1973). The fifth and sixth formant generators, and the multiplier had only the noise source as an excitation source. The multiplier was used to simulate the broadband noise for sounds that did not have any defined resonance structure. The radiation load was simulated by the first difference of the combined output of the cascade and the parallel filter banks. The glottal source model in this synthesizer generated two types of pulses: one for synthesizing vowels, nasals, etc., and the other for synthesizing mixed sounds and the "voice-bar." A cascade of a second order resonator and a second order anti-resonator was used to obtain glottal source pulses with -12 dB/oct spectral tilt for vowels, etc. The other glottal source model was a cascade of two second order resonators to obtain glottal source pulses with -24 dB/oct spectral tilts for mixed sounds and voice-bar. The noise source was a built-in random number generator with white-noise characteristics in order to simulate the flat spectrum of the air pressure at the lungs. The noise source was filtered by a first order lowpass filter to simulate the conversion of the constant pressure source at the lungs to the source volume-velocity at the glottal constriction. This synthesizer was capable of synthesizing highly intelligible speech.

Holmes (1983) proposed that the parallel configuration of the filter bank was superior to the cascade configuration of the filter bank even for the synthesis of vowels. The author argued that the "all-pole" cascade configuration was not theoretically correct for vowel production, since the assumption of plane wave propagation does not hold above 3 KHz, in contrast with the original assumption of 8 KHz; furthermore, there were notches ("zeros") present in the vowel spectrum above 3 KHz, which the cascade configuration of resonators could not simulate. Other argument was that there was no obvious reason why the "pole" frequencies of the filter bank in this region should be the same as the resonant modes of the vocal tract. The "critical-bands" in

this frequency range were of the order of 500 Hz wide and so the spectral fine details in this frequency range were not perceptually important. The transfer function of the cascade filter bank, which was a product of the transfer function of each resonator could be represented as the sum of the transfer function of the resonators by taking partial fractions. It was shown that simulating the cascade filter bank by the parallel filter bank using this method leads to problems in the low frequency region of the magnitude frequency response, when the amplitude control parameter of any one resonator is slightly changed. A new configuration was proposed where each parallel filter bank resonator did not have a significant skirt response (response beyond the specified bandwidth of the resonator.) With this configuration, it was possible to change the amplitude of a formant without affecting the amplitudes of other formants. A phase correction network was added to the first formant generator in order to obtain the correct values of the magnitude frequency response in the low frequency range. The first formant generator shared its "normal frequency range" with a nasal resonator having a fixed center frequency and bandwidth. A single amplitude control parameter could specify the gains of both the first formant and the nasal resonator. The fourth formant generator was a broadband filter bank consisting of three resonators in cascade. The fourth formant amplitude control parameter actually specified the amplitude level of the magnitude frequency response of the filter bank at the high frequencies. The output of each formant generator (resonator) was added after alternate scaling by +1 and -1 in order to prevent creation of "notches" in the magnitude frequency response. This synthesizer could synthesize highly intelligible speech.

Verhelst and Nilens, (1986) described a "modified super-position speech synthesizer" which consisted of two cascade filter banks arranged in parallel to create a cascade branch. This configuration reduced the transients generated in the filter banks when the formant parameters change drastically at the speech frame



boundaries. At the beginning of each frame, the coefficients of the filters in the "old" filter bank were updated according to the formant parameters for the new frame and the "old" filter bank's memory was cleared. The excitation source was connected only to this filter bank and was disconnected from the other filter bank. The filter bank updated for the previous frame was left "free-running," i.e., producing output only from the stored energy. In this algorithm, the energy in the updated filter bank increases while the energy in the other filter bank is dissipated in the new frame. The output from the two cascade filter banks was added to obtain a smooth speech waveform.

Pinto, et al. (1989) described a synthesizer which was an outgrowth of Klatt's cascade/parallel formant synthesizer [Klatt, 1980]. This synthesizer used the LF (Liljencrants and Fant) model (Fant et al., 1985) and Ananthapadmanabha's and Fant's (1982) circuit as the two optional glottal source models. The first glottal source model added flexibility in terms of creating various shapes of the glottal source pulses and the second glottal source model added an interface to the physiological models of the vocal folds.

Klatt and Klatt (1990) described modification to the cascade/parallel configuration of the formant synthesizer described earlier in Klatt (1980). The modified synthesizer incorporated three glottal source models: impulse train as a glottal source model, a modified Klatt's glottal source model and the modified LF model [Fant et al., 1985]. Klatt's modified glottal source model had seven parameters to control both the time and frequency domain characteristics of the glottal source pulses. By controlling the time and frequency domain features related to the vibratory patterns of the vocal folds, the modified Klatt's model could successfully mimic the 'temporary laryngealized offset accompanied by double pulsing' and 'breathy falling-pitch offset accompanied at the end of an utterance.' The authors suggested that the naturalness of the synthetic speech can be increased by adding "flutter" (slowly



varying statical fluctuations to the pitch-period) and “double pulsing” (reduced amplitude of the glottal source pulses in the alternate periods). But the exact perceptual correlations of these two features were not completely studied. The modified LF model, provided in this synthesizer, has been proven to be very useful in synthesizing natural sounding normal speech and various speech disorders [Fant et al., 1985; Gobl, 1989; Lee and Childers, 1989]. This glottal source model is discussed in detail in Chapters 4 and 5.

The cascade/parallel synthesizer configuration used the cascade filter bank to synthesize voiced and aspirated sounds. The parallel filter bank (also common to the all-parallel synthesizer configuration) was used to synthesize fricatives. The all-parallel synthesizer configuration used two separate parallel filter banks: one for synthesizing the voiced and aspirated sounds and other for synthesizing the fricatives. In the early synthesizer [Klatt, 1980] one and the same parallel filter bank was used for synthesizing voiced, and aspirated sounds and fricatives. The most significant tracheal formant and anti-formant pair was simulated in the manner similar to the simulation of nasal formant and anti-formant pair. This synthesizer could copy utterances from several female and male speakers with very good perceptual fidelity. The requirement for such fidelity was a good initial match of fundamental frequency contour and the short-time spectra sampled throughout an utterance. The value of the first formant frequency and bandwidth could be dynamically changed during the open phase of the glottal source pulse to simulate rapid changes in the glottal losses as the vocal folds opened and closed.

Holmes et al. (1990) modified their all-parallel formant synthesizer to extend the bandwidth of synthesized speech from 4 KHz to 8 KHz. The previously implemented fourth formant filter bank was replaced by a broadband filter bank consisting of four bandpass filters to completely cover the frequency range from above 3 KHz to 8 KHz. This filter bank was considered to be sufficient for the “normal fourth formant and

above frequency range." The authors argued that, in this frequency range, as long as the signal level at high frequencies was appropriate, the spectral fine details were not as important. They obtained high-quality synthesis for both male and female speech when the input parameters were carefully controlled.

### Limitations of the Current Formant Synthesizers

At present the two most popular formant synthesizers are: 1) the cascade/parallel formant synthesizer developed by Klatt [Klatt, 1980] and 2) the parallel formant synthesizer developed by Holmes [Holmes, 1983]. Under separate testing conditions, the software for both formant synthesizers demonstrated a capability to synthesize high-quality natural sounding speech when the parameters of the synthesizers were carefully controlled. For each synthesizer, the user could specify a variable number of parameters from a fixed list of parameters. The other parameters (not specified by the user) in the parameters list were assigned default values at the start-up (initiation of synthesis). Some of the parameters, such as formant frequencies, bandwidths and amplitudes, were also specified as time-varying parameters (changing for each frame of the utterance being synthesized).

While a certain flexibility existed in specifying the parameters of the synthesizers, there was no flexibility to change the configuration (architecture) of either of these synthesizers. For example, these synthesizers lacked the flexibility to provide the user with several glottal source models, simplicity in the procedures to specify glottal source models' parameters and to change the number of formants used for the synthesis of a particular utterance. The synthesizer parameter specification procedure should allow for specifying a variable number of resonators and the synthesizer architecture should be flexible enough to synthesize a smooth output waveform from the specified number of resonators (formants).

### Limitations of the Source Models

Klatt's (1980) synthesizer has a simple glottal source model that generates glottal pulses with a spectral tilt of  $-12$  dB/oct or  $-24$  dB/oct. The shape of the pulses generated by this glottal source model could not be controlled, a drawback for conducting various experiments. A fixed (built-in) white-noise source was used to synthesize aspirated sounds and fricatives. The modified cascade/parallel synthesizer by Klatt [Klatt and Klatt, 1990] provided the user with three glottal source models: impulse train generator, modified Klatt's model and modified LF model. The Klatt modified glottal source model is flexible to the extent that it can generate glottal source pulses of various shapes. Also, period to period variations in the peak amplitude and pitch period of the glottal source pulses could be brought about to introduce "shimmer" and "jitter" in the synthesized speech. But the rules to obtain the values of the model parameters for generating the glottal source pulses with the desired time and frequency domain characteristics were not available. The modified LF model is known to be highly flexible for generating glottal source pulses of various shapes. In the modified synthesizer, separate noise sources for synthesizing aspirated and friction sounds were provided. But both noise sources were built-in random number generators. These random number generators generated the same random number sequences at every instance of synthesis, since the seed values of the random number generators could not be changed. Some listening tests, e.g., determining the intelligibility of a certain synthesized consonant, may require several tokens of the same consonant to be synthesized with several different random number sequences. There was no provision to use external sampled data waveforms, which may have been obtained from the analysis of natural speech, as glottal source, aspiration noise source or friction noise source.

Holmes' (1983) all-parallel formant synthesizer did not use a glottal source model to generate stylized glottal source pulses. The excitation source was obtained by applying the spectral flattening process [Holmes, 1983] to the speech of a typical talker. The output of the parallel filter bank was passed through a fixed lowpass filter to obtain the -6 dB/oct spectral tilt above 640 Hz in the synthesized speech. A single noise source (built-in random number generator) was used for synthesizing both aspirated sounds and fricatives. This synthesizer used the amplitude control parameters to control not only the amplitudes of the resonances of the vocal tract but also to control those aspects of the speech spectrum that were normally controlled by the glottal source model in other synthesizers. Clearly, such an architecture is not useful for conducting a wide variety of experiments with synthetic speech.

#### Limitation in the Filter Bank Configuration

The rigidity in the configuration of filter banks in the modified Klatt's cascade/parallel formant synthesizer [Klatt and Klatt, 1990] has not been reduced compared to its earlier version described in Klatt (1980). Same is the case with the recent version of Holmes' all-parallel formant synthesizer [Holmes et al., 1990]. The two possible configurations of the Klatt synthesizer configuration are the cascade/parallel synthesizer configuration and an all-parallel synthesizer configuration. In Klatt's cascade/parallel synthesizer configuration, the voiced sounds were synthesized by exciting the cascade filter bank and the unvoiced sounds were synthesized by exciting the parallel filter bank. In the all-parallel synthesizer configuration and also in Holmes' all-parallel synthesizer, all sounds were synthesized by the parallel filter bank. The cascade filter bank had a fixed number of resonators and anti-resonators. The parallel filter bank had a fixed number of resonators in each filter bank. The formants and the anti-formants were assigned to the fixed (specific) resonators and the anti-resonators in the filter banks and this assignment remained

fixed until the completion of the synthesis of an utterance. In Klatt's cascade/parallel formant synthesizer there exists some flexibility in the configuration of the filter banks for synthesizing female voices and nasal sounds. When synthesizing a female voice the number of resonators in the cascade branch could be reduced by one (by removing the fifth formant resonator) at the start-up. But, the resonators and the anti-resonators in the cascade filter bank and/or the parallel filter bank could not be removed during the synthesis. When synthesizing nasals, a nasal resonator and anti-resonator pair could be added in series with the cascade filter bank. A nasal resonator could also be effectively added to the parallel filter bank. But the same specifications of center frequencies and bandwidths for the resonator and anti-resonator pair were used for all the nasals. In Holmes' all-parallel formant synthesizer, the shape of the magnitude frequency response of the parallel filter bank at the high frequencies was fixed, since the shape of the magnitude frequency response of the broadband filter bank is fixed. Therefore, there is no control over the shape of the magnitude frequency response of the parallel filter bank at the high frequencies, except for the overall amplitude level. In the parallel filter bank of both synthesizers, any resonator can be "removed," indirectly, by reducing the gain of a resonator to a very small value (i.e., by reducing the value of the amplitude control parameter of that resonator). Then the contribution of that filter to the total output of the parallel filter bank reduces to a negligible amount. But to achieve this, very precise control of the formant amplitude tracks was required. Also, the precise specification of the formant tracks of the temporarily "unused" resonators was required. In both synthesizers, the default values of the formant frequencies and bandwidths were automatically assigned to the center frequencies and bandwidths of the resonators not specified by the user with no option given to the user.

## APPENDIX B EXCITATION SOURCE: GLOTTAL AND NOISE

### Glottal Source Models

The parametric and non-parametric glottal source models incorporated in the flexible formant synthesizer are described in the following sub-sections. The block diagrams and the typical waveforms of these glottal source models are given in Figure B-1 through Figure B-5. Each parametric glottal source model has a list of parameters to specify various characteristics (shape) of the glottal source pulses. The list of parameters for each non-parametric glottal source model consists of the amplitude related parameters. The list of parameters for the glottal source models are given in Table B-I to Table B-V. These tables also give a brief description of the parameters of the glottal source models.

### Parametric Glottal Source Models

The following parametric glottal source models are incorporated in the flexible formant synthesizer.

1) Klatt's Model [Klatt, 1980]: This source model, described in Klatt (1980), generates glottal source pulses when excited by an impulse train. This model consists of a second order resonator in series with a parallel combination of another second order resonator and an anti-resonator. The series combination of resonator and anti-resonator is used when synthesizing vowels, nasal, etc. The series combination of two resonators is used when synthesizing a voice-bar, voiced plosives and

fricatives, and breathy sounding speech. This model is primitive in a sense because it simulates only the overall frequency domain characteristics of typical glottal flow pulses. This model cannot simulate the time domain characteristics of glottal flow pulses, such as right-skewness, duty cycle, etc. Klatt and Klatt (1990) has described another source model (explained in Appendix A) that can simulate both the frequency and time domain characteristics of typical glottal flow pulses. However, the rules to manipulate the parameters of this glottal source model, in order to obtain various shapes of the glottal flow pulses, are very primitive. The block diagram and the typical waveforms of this glottal source model are shown in Figure B-1, and the list of the parameters is given in Table B-I. The glottal source model described in Klatt (1980) is included in the flexible formant synthesizer only for historical purposes.

2) Two/Three Pole Model [Lee, 1988]: This glottal source model is implemented as a cascade of three first order IIR (Infinite Impulse Response) digital filters, each representing a single positive “real pole” inside the unit circle in the  $z$ -plane. The glottal source pulses are generated when the cascade of filters is excited by an impulse train. Each IIR (Infinite Impulse Response) filter is capable of providing  $-6\text{dB/oct}$  additional slope to the spectrum of a waveform passing through it. Like Klatt’s glottal source model [Klatt, 1980], this model is primitive because the model simulates only the overall frequency domain characteristics of typical glottal flow pulses. This model cannot simulate the time domain characteristics of glottal flow pulses, such as right-skewness, duty cycle, etc. Markel and Gray (1976) proposed the “two-pole” model for synthesizing “normal” sounding speech. Lee (1988) has modelled the spectral-tilt of the glottal flow pulses for breathy and falsetto speech sounds with the “three-pole” model. The block diagram and the typical waveforms of this glottal source model are shown in Figure B-2, and the list of the parameters is given in Table B-II.



3) LF Model (Liljencrants and Fant) [Fant et al., 1985]: The parameters of this model specify the derivative of the glottal source pulses and the time-integral of these pulses gives the glottal source pulses. The time domain features of the glottal flow pulses, such as the negative peak and positive peak of the waveform, right skewness, duty-cycle, etc., can be simulated with the parameters of this model. The frequency domain features of the glottal flow pulses, such as spectral tilt, location(s) of spectral "zeros," intensity of the fundamental component, etc. can be controlled by the parameters of this model. The LF model is good for non-interactive flow parameterization in the sense that it can fully ensure an overall fit to commonly encountered glottal pulse shapes, has a minimal number of parameters, and is flexible in its ability to match extreme phonations [Klatt and Klatt, 1990]. The typical waveforms of this glottal source model are shown in Figure B-3 and the list of the parameters is given in Table B-III. This glottal source model is described in detail in chapters 4 and 5.

4) New Glottal Source Model [Lalwani and Childers, 1991b]: This glottal source model uses the LF model to generate glottal source pulses. The LF model is used in conjunction with the aspiration source generator and the pitch-period perturbation generator to incorporate modulated aspiration noise, period-to-period variations in the pitch-period (jitter) and peak amplitude (shimmer) of glottal source pulses. This model has been designed to synthesize high-quality "normal" speech and also various speech disorders. The block diagram of this glottal source model is shown in Figure B-4, and the list of the parameters is given in Table B-IV. This model is described in detail in chapters 4 and 5.

5) Ananthapadmanabha's and Fant's Circuit [Ananthapadmanabha and Fant, 1982]: This model simulates the aerodynamic characteristics of the sound (source) production process in the human speech production system. The lung pressure, the glottal volume velocity, the sub-glottal acoustic impedance, the glottal acoustic



impedence (which is inversely proportional to the time varying glottal area) and the supra-glottal acoustic impedance are simulated by an analog (equivalent) electric circuit. This model simulates source tract interaction that is present in the human speech production system when the glottal-open area is large. When the glottal-open area is large, the source cannot be considered to be independent of (de-coupled from) the vocal-tract acoustic load. The glottal source is loaded by the acoustical impedance of the vocal tract. Also, the sub-glottal system affects the resonance characteristics of the vocal tract. In this model, a single sub-glottal formant and the first formant of the vocal tract are assumed to affect the glottal source generation. The effect of source-tract interaction is presumed to cause right-skewing of the glottal flow pulses and also produce first formant ripple super-imposed on the glottal source pulses [Fant and Ananthapadmanabha, 1982]. This model produces glottal source pulses that have right-skewing and super-imposed formant ripple. The glottal area waveform can either be an arbitrary glottal area function or the waveform generated by some model such as Titze's area function model [Titze, 1984]. The standard values for the resistors, inductors and capacitors for modeling five vowels are given in Ananthapadmanabha and Fant, (1982) and a method to interpolate their values for the other vowels is described in Pinto (1987). This model can be used to test the physiological models that simulate glottal area waveforms for various vocal fold configurations. This model will be implemented in the flexible formant synthesizer. Temporarily, the user can generate waveforms of glottal source pulses using other implementations of this model (external to the flexible formant synthesizer) and then use these waveforms as a non-parametric glottal source model in the flexible formant synthesizer.

### Non-parametric Glottal Source Models

The non-parametric glottal source models are classified depending upon the number of pulses in the glottal source waveform used as a glottal source. The following non-parametric models can be specified in the flexible formant synthesizer.

- 1) **Impulse Train Generator:** When this model is specified as a glottal source, impulses are generated at the rate specified by the fundamental frequency parameter. The amplitude of each impulse is specified by the voicing gain parameter. This is a very primitive glottal source model because it does not simulate either the time-domain or the frequency domain characteristics of typical glottal flow pulses. This model is useful for obtaining the impulse response of the filter banks. A typical impulse train waveform is shown in Figure B-5a, and the list of the parameters for this glottal source model is give in Table B-V.
- 2) **Single-Pulse Waveform:** A single pulse of an inverse filtered speech waveform or a stylized pulse waveform generated by some parametric model can be used as a glottal source when this model is selected. This single pulse waveform is repeated at the rate specified by the fundamental frequency parameter to generate glottal source. This model uses a time-warping algorithm to match the number of samples in the input glottal pulse with the number of samples in the pitch-period (specified by fundamental frequency parameter and the sampling rate). The pitch period is defined as the interval of repetition of the glottal source pulses and is the multiplicative inverse of the value of the fundamental frequency. A typical glottal source pulse waveform is shown in Figure B-5b, and the list of the parameters for this glottal source model is given in Table B-V.
- 3) **Multi-Pulse Waveform:** A multi-pulse inverse filtered speech waveform obtained from an all voiced sentence or a sequence of stylized pulses generated by some parametric model can be used as the glottal source. The pitch-period is determined by

the rate of repetition of the pulses in the glottal source waveform. The pitch-period of the pulses in the glottal source waveform is not analyzed by the flexible formant synthesizer. Therefore, the user does not have explicit control over the pitch-period of the glottal source pulses (i.e., the specification of the fundamental frequency contour is ineffective in this case). Also, the shape of the pulses in the glottal source waveform cannot be modified. The time-warping algorithm to match the pitch-period of the pulses in the glottal source waveform with the pitch-period specified by the fundamental frequency contour cannot be used. Normally, the fixed-frame synthesis method is employed when this model is used as a glottal source. If pitch-synchronous synthesis has to be employed, the flag "PITCH\_SYNC" should be set. In this case, care should be taken to match the pitch-period of the successive pulses in the glottal source waveform. The synthesis of an utterance is terminated when the number of synthesized samples is equal to the number of samples of the glottal source waveform (even if it is not the specified number of samples to be synthesized). A typical glottal source pulse waveform is shown in Figure B-5b, and the list of the parameters for this glottal source model is given in Table B-V.

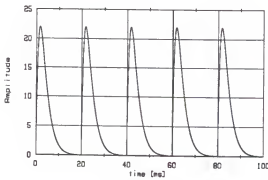
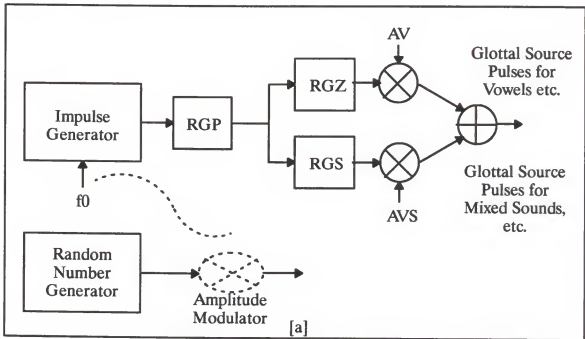
### Noise Sources

The following noise source models are used in the flexible formant synthesizer.

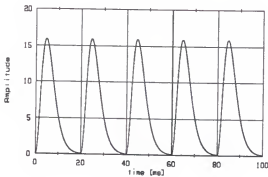
- 1) Built-in Random Number Generator: High-level programming languages such as "C" and "FORTRAN77" provide math-libraries that have "functions" that return random numbers. By repeatedly invoking a random number function, a random number sequence with a uniform distribution within the range  $\pm 0.5$  and with white-noise characteristics can be generated. This random number sequence is used as a noise source in the software implementation. A new random number sequence with pseudo-Gaussian distribution (Gaussian-like distribution within

limited range) limited to the range of  $\pm 0.5$  is created from the average values of the 16 consecutive random numbers in the original random number sequence with uniform distribution (central limit theorem). The random number sequence is then scaled by an appropriate scale factor to obtain a unit power random number sequence. The block diagram and the typical waveforms of this noise source model are shown in Figure B-6. The noise source may also be amplitude-modulated and filtered as shown in the block diagram.

2) External Random Number Tables: The external random number sequences stored in files can also be used as a noise source. Different random number sequences can be generated either from the same random number function with different seed (starting) values or from different random number generators. The error signal sequences obtained during LPC analysis of unvoiced sounds, can also be used as a noise source. If the length of the external random number sequence is smaller than the number of samples to be synthesized, the built-in random number generator is automatically "turned-on" to append the random numbers to the original random number sequence. There is no provision to change the distribution, range or the mean value of the external random number sequence. But, the external random number sequence can be amplitude-modulated and/or filtered as shown in the block diagram in Figure B-6a.



[b]



[c]

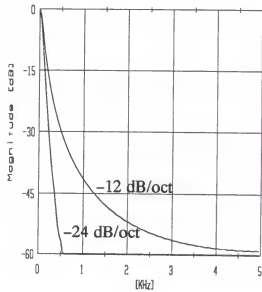
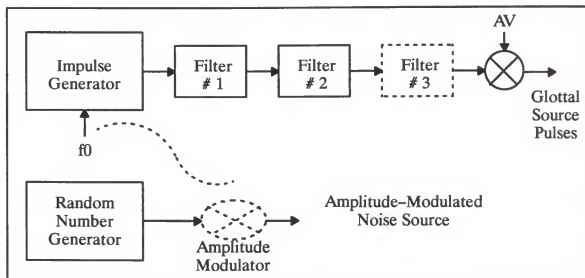
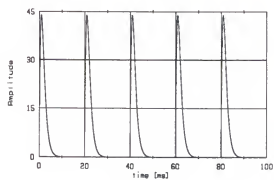


Figure B-1: Klatt's Model

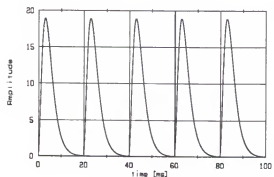
- Simple block diagram for Klatt's Model
- Glottal source pulses for vowels, etc.
- Glottal source pulses for mixed sounds, etc.
- Envelopes of the spectra of the two waveforms



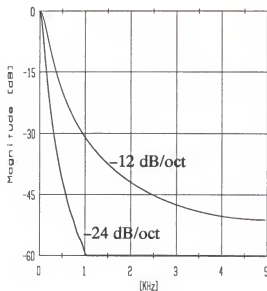
[a]



[b]



[c]



[d]

Figure B-2: Two/three Pole Model

- Simple block diagram for Two/Three Pole Model
- Glottal source pulses for vowels, etc.
- Glottal source pulses for mixed sounds, etc.
- Envelopes of the spectra of the two waveforms

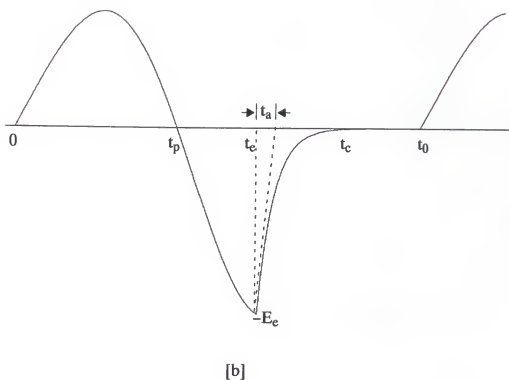
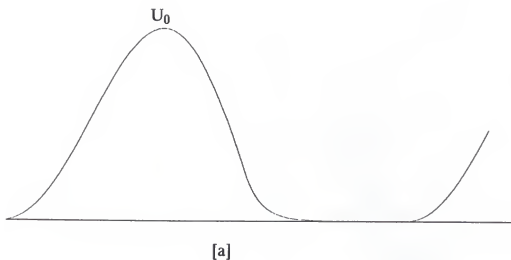


Figure B-3: LF model time function  
 a) Integrated LF model time function (glottal source pulse)  
 b) LF model time function (differentiated glottal source pulse)

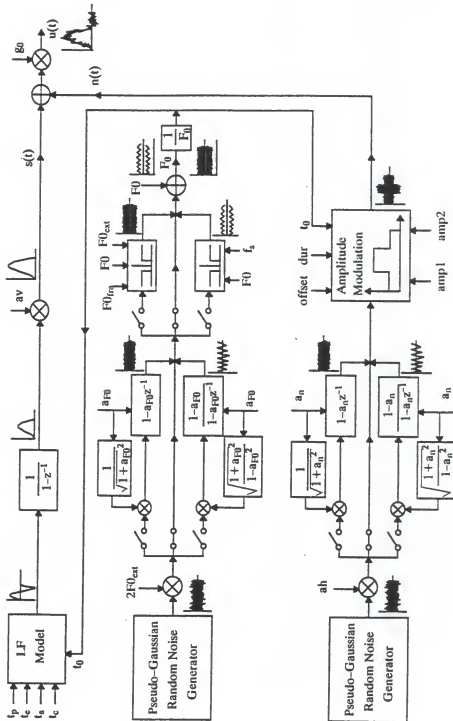


Figure B-4: New Glottal Source Model



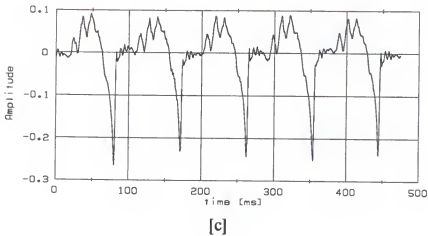
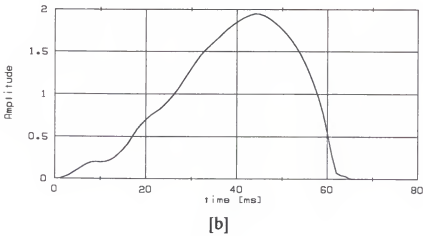
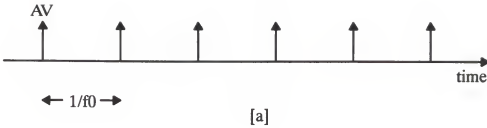
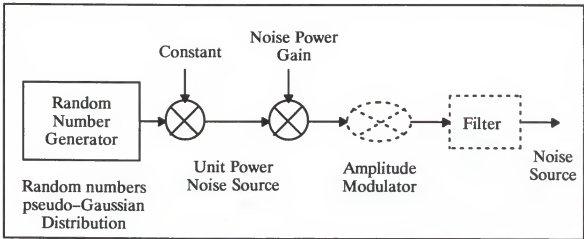
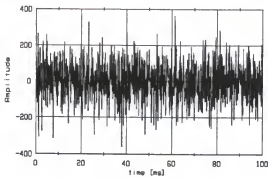


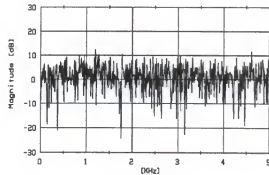
Figure B-5: Non-parametric glottal source models  
 a) Impulse train as a glottal source model  
 b) Single pulse of inverse filtered speech waveform as a Single-pulse glottal source model  
 c) Multiple pulses of the differentiated inverse filtered speech waveform as a Multi-pulse glottal source model



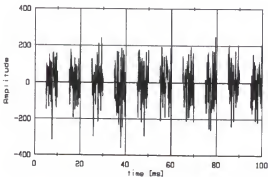
[a]



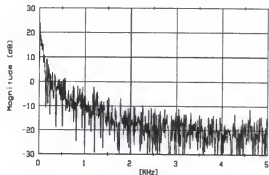
[b]



[c]



[d]



[e]

Figure B-6: Noise Source Model

- a) Block diagram for Noise Source Model
- b) White-noise source
- c) Spectrum of white-noise source
- d) Amplitude-modulated noise source
- e) Spectrum of lowpass filtered noise source

Table B-I

Table of parameters for the Klatt's Model  
with minimum, typical and maximum values

| # Parameter Name                                                    | Min. Value | Typical Value | Max. Value |
|---------------------------------------------------------------------|------------|---------------|------------|
| # Glottal source gain when synthesizing mixed sounds, etc. (in dB)  |            |               |            |
| 1) avs                                                              | 0.0        | 60.0          | 60.0       |
| # Gain of the common resonator, RGP (in dB)                         |            |               |            |
| 2) a1                                                               | 0.0        | 0.0           | 80.0       |
| # Bandwidth of the common resonator, RGP (in Hz)                    |            |               |            |
| 3) b1                                                               | 100.0      | 100.0         | 2000.0     |
| # Center frequency of the common resonator, RGP (in Hz)             |            |               |            |
| 4) f1                                                               | 0.0        | 0.0           | 60.0       |
| # Gain of the resonator, RGS (in dB)                                |            |               |            |
| 5) a2                                                               | 0.0        | 0.0           | 80.0       |
| # Bandwidth of the resonator, RGS (in Hz)                           |            |               |            |
| 6) b2                                                               | 100.0      | 100.0         | 2000.0     |
| # Center frequency of the resonator, RGS (in Hz)                    |            |               |            |
| 7) f2                                                               | 0.0        | 0.0           | 60.0       |
| # Gain of the anti-resonator, RGZ (in dB)                           |            |               |            |
| 8) a3                                                               | 0.0        | 0.0           | 80.0       |
| # Bandwidth of the anti-resonator, RGZ (in Hz)                      |            |               |            |
| 9) b3                                                               | 100.0      | 100.0         | 2000.0     |
| # Center frequency of the anti-resonator, RGZ (in Hz)               |            |               |            |
| 10) f3                                                              | 0.0        | 0.0           | 60.0       |
| # Parameters for amplitude-modulation of noise source               |            |               |            |
| # Duration of the first part of the amplitude-modulation waveform   |            |               |            |
| # specified as a fraction of pitch-period.                          |            |               |            |
| 11) offset                                                          | 0.0        | 0.5           | 1.0        |
| # Duration of the second part of the amplitude-modulation waveform  |            |               |            |
| # specified as a fraction of pitch-period.                          |            |               |            |
| 12) dur                                                             | 0.0        | 0.5           | 1.0        |
| # Amplitude of the first part of the amplitude-modulation waveform  |            |               |            |
| 13) amp1                                                            | 0.0        | 0.5           | 1.0        |
| # Amplitude of the second part of the amplitude-modulation waveform |            |               |            |
| 14) amp2                                                            | 0.0        | 0.5           | 1.0        |
| #                                                                   |            |               |            |

Table B-I ... Continued

| # Parameter Name                                                                        | Min. Value | Typical Value | Max. Value |
|-----------------------------------------------------------------------------------------|------------|---------------|------------|
| # Choices for interpreting the values of the gain parameters "av" and "avs"             |            |               |            |
| # 1 -> total power in the glottal source pulse                                          |            |               |            |
| # 2 -> total energy in the glottal source pulse                                         |            |               |            |
| # 3 -> peak amplitude of the glottal source pulse                                       |            |               |            |
| # 4 -> peak negative amplitude of the differentiated glottal source pulse               |            |               |            |
| # 5 -> cannot be used for this model                                                    |            |               |            |
| # 6 -> these parameters have no effect on the glottal source pulse                      |            |               |            |
| #                                                                                       |            |               |            |
| 15) typ_gain                                                                            | 1          | 1             | 6          |
| # Scale factor for controlling the amplitude of the glottal source pulse without        |            |               |            |
| # changing the values of the "av" and "avs" parameters                                  |            |               |            |
| 16) scale                                                                               | 0.0        | 1.0           | 1000.0     |
| # Flag to indicate the simulation of perceptual effect of varying "stress" with varying |            |               |            |
| # "pitch." When set, the glottal source pulse is scaled by the value of fundamental     |            |               |            |
| # frequency parameter.                                                                  |            |               |            |
| 17) F0                                                                                  | 0          | 1             | 1          |
| # Flag to indicate initialization of the filters in the glottal source model before the |            |               |            |
| # onset of following glottal source pulse                                               |            |               |            |
| 18) INI                                                                                 | 0          | 1             | 1          |
| #                                                                                       |            |               |            |

Table B-II

Table of parameters for the Two/Three Pole Model  
with minimum, typical and maximum values

| # Parameter Name                                                                        | Min. Value | Typical Value | Max. Value |
|-----------------------------------------------------------------------------------------|------------|---------------|------------|
| # Filter coefficient for the first order filter # 1                                     |            |               |            |
| 1) coeff1                                                                               | 0.0        | 0.99          | 0.99       |
| # Filter coefficient for the first order filter # 2                                     |            |               |            |
| 2) coeff2                                                                               | 0.0        | 0.99          | 0.99       |
| # Filter coefficient for the first order filter # 3                                     |            |               |            |
| 3) coeff3                                                                               | 0.0        | 0.99          | 0.99       |
| # Parameters for amplitude-modulation of noise source                                   |            |               |            |
| # Duration of the first part of the amplitude-modulation waveform                       |            |               |            |
| # specified as a fraction of pitch-period.                                              |            |               |            |
| 4) offset                                                                               | 0.0        | 0.5           | 1.0        |
| # Duration of the second part of the amplitude-modulation waveform                      |            |               |            |
| # specified as a fraction of pitch-period.                                              |            |               |            |
| 5) dur                                                                                  | 0.0        | 0.5           | 1.0        |
| # Amplitude of the first part of the amplitude-modulation waveform                      |            |               |            |
| 6) amp1                                                                                 | 0.0        | 0.5           | 1.0        |
| # Amplitude of the second part of the amplitude-modulation waveform                     |            |               |            |
| 7) amp2                                                                                 | 0.0        | 0.5           | 1.0        |
| # Choices for interpreting the values of the gain parameter "av"                        |            |               |            |
| # 1 -> total power in the glottal source pulse                                          |            |               |            |
| # 2 -> total energy in the glottal source pulse                                         |            |               |            |
| # 3 -> peak amplitude of the glottal source pulse                                       |            |               |            |
| # 4 -> peak negative amplitude of the differentiated glottal source pulse               |            |               |            |
| # 5 -> cannot be used for this model                                                    |            |               |            |
| # 6 -> these parameters have no effect on the glottal source pulse                      |            |               |            |
| 8) typ_gain                                                                             | 1          | 1             | 6          |
| # Scale factor for controlling the amplitude of the glottal source pulse without        |            |               |            |
| # changing the values of the "av" parameter                                             |            |               |            |
| 9) scale                                                                                | 0.0        | 1.0           | 1000.0     |
| # Flag to indicate the simulation of perceptual effect of varying "stress" with varying |            |               |            |
| # "pitch." When set, the glottal source pulse is scaled by the value of fundamental     |            |               |            |
| # frequency parameter.                                                                  |            |               |            |
| 10) F0                                                                                  | 0          | 1             | 1          |
| # Flag to indicate initialization of the filters in the glottal source model before the |            |               |            |
| # onset of following glottal source pulse                                               |            |               |            |
| 11) INI                                                                                 | 0          | 1             | 1          |

Table B-III

Table of parameters for the LF Model  
with minimum, typical and maximum values

| # Parameter Name                                                                                                 | Min. Value | Typical Value | Max. Value |
|------------------------------------------------------------------------------------------------------------------|------------|---------------|------------|
| # LF model's direct synthesis parameters                                                                         |            |               |            |
| # Peak negative amplitude of the differentiated glottal source pulse                                             |            |               |            |
| 1) ee                                                                                                            | 0.0        | 50.0          | 100.0      |
| # Growth constant of the exponentially increasing sinusoid (L model)                                             |            |               |            |
| 2) alpha                                                                                                         | 0.0        | 0.05          | 1.0        |
| # Fundamental frequency of the exponentially increasing sinusoid (L model)                                       |            |               |            |
| 3) wg                                                                                                            | 0.0        | 100.0         | 100.0      |
| # Decay constant of the exponentially decaying recovery phase (recovery phase)                                   |            |               |            |
| 4) eps                                                                                                           | 0.0        | 0.1           | 10.0       |
| # LF model's timing parameters                                                                                   |            |               |            |
| # Instant of peak in the glottal source pulse specified as % of pitch-period                                     |            |               |            |
| 5) tp                                                                                                            | 0.0        | 50.0          | 99.0       |
| # Instant of the negative peak in the differentiated glottal source pulse<br># specified as % of pitch-period    |            |               |            |
| 6) te                                                                                                            | 0.0        | 60.0          | 100.0      |
| # Time constant of the recovery phase as % of pitch-period                                                       |            |               |            |
| 7) ta                                                                                                            | 0.0        | 60.0          | 100.0      |
| # Choice for specifying the LF model's parameters                                                                |            |               |            |
| # 1 -> direct synthesis parameters                                                                               |            |               |            |
| # 2 -> timing parameters                                                                                         |            |               |            |
| 8) par                                                                                                           | 1          | 2             | 2          |
| # Parameters for amplitude-modulation of noise source                                                            |            |               |            |
| # Duration of the first part of the amplitude-modulation waveform<br># specified as a fraction of pitch-period.  |            |               |            |
| 11) offset                                                                                                       | 0.0        | 0.5           | 1.0        |
| # Duration of the second part of the amplitude-modulation waveform<br># specified as a fraction of pitch-period. |            |               |            |
| 12) dur                                                                                                          | 0.0        | 0.5           | 1.0        |
| # Amplitude of the first part of the amplitude-modulation waveform                                               |            |               |            |
| 13) amp1                                                                                                         | 0.0        | 0.5           | 1.0        |
| # Amplitude of the second part of the amplitude-modulation waveform                                              |            |               |            |
| 14) amp2                                                                                                         | 0.0        | 0.5           | 1.0        |
| # Choices for interpreting the values of the gain parameter "av"                                                 |            |               |            |
| #                                                                                                                |            |               |            |

Table B-III ... Continued

| # Parameter Name                                                                        | Min. Value | Typical Value | Max. Value |
|-----------------------------------------------------------------------------------------|------------|---------------|------------|
| # 1 -> total power in the glottal source pulse                                          |            |               |            |
| # 2 -> total energy in the glottal source pulse                                         |            |               |            |
| # 3 -> peak amplitude of the glottal source pulse                                       |            |               |            |
| # 4 -> peak negative amplitude of the differentiated glottal source pulse               |            |               |            |
| # 5 -> cannot be used for this model                                                    |            |               |            |
| # 6 -> these parameters have no effect on the glottal source pulse                      |            |               |            |
| #                                                                                       |            |               |            |
| 15) typ_gain                                                                            | 1          | 1             | 6          |
| # Scale factor for controlling the amplitude of the glottal source pulse without        |            |               |            |
| # changing the values of the "av" parameter                                             |            |               |            |
| 16) scale                                                                               | 0.0        | 1.0           | 1000.0     |
| # Flag to indicate the simulation of perceptual effect of varying "stress" with varying |            |               |            |
| # "pitch." When set, the glottal source pulse is scaled by the value of fundamental     |            |               |            |
| # frequency parameter.                                                                  |            |               |            |
| 17) F0                                                                                  | 0          | 1             | 1          |
| # Flag to indicate initialization of the filters in the glottal source model before the |            |               |            |
| # onset of following glottal source pulse                                               |            |               |            |
| 18) INI                                                                                 | 0          | 1             | 1          |
| #                                                                                       |            |               |            |

Table B-IV

Table of parameters for the New Glottal Source Model  
with minimum, typical and maximum values

| # Parameter Name                                                                                                                                                                                             | Min. Value | Typical Value | Max. Value |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|---------------|------------|
| # LF model's direct synthesis parameters                                                                                                                                                                     |            |               |            |
| # Peak negative amplitude of the differentiated glottal source pulse                                                                                                                                         |            |               |            |
| 1) ee                                                                                                                                                                                                        | 0.0        | 50.0          | 100.0      |
| # Growth constant of the exponentially increasing sinusoid (L model)                                                                                                                                         |            |               |            |
| 2) alpha                                                                                                                                                                                                     | 0.0        | 0.05          | 1.0        |
| # Fundamental frequency of the exponentially increasing sinusoid (L model)                                                                                                                                   |            |               |            |
| 3) wg                                                                                                                                                                                                        | 0.0        | 100.0         | 100.0      |
| # Decay constant of the exponentially decaying recovery phase (recovery phase)                                                                                                                               |            |               |            |
| 4) eps                                                                                                                                                                                                       | 0.0        | 0.1           | 10.0       |
| # LF model's timing parameters                                                                                                                                                                               |            |               |            |
| # Instant of peak in the glottal source pulse specified as % of pitch-period                                                                                                                                 |            |               |            |
| 5) tp                                                                                                                                                                                                        | 0.0        | 50.0          | 99.0       |
| # Instant of the negative peak in the differentiated glottal source pulse<br># specified as % of pitch-period                                                                                                |            |               |            |
| 6) te                                                                                                                                                                                                        | 0.0        | 60.0          | 100.0      |
| # Instant of the closure of glottal source pulse specified as % of pitch-period                                                                                                                              |            |               |            |
| 7) tc                                                                                                                                                                                                        | 0.0        | 60.0          | 100.0      |
| # Time constant of the recovery phase as % of pitch-period                                                                                                                                                   |            |               |            |
| 8) ta                                                                                                                                                                                                        | 0.0        | 60.0          | 100.0      |
| # Choice for specifying the LF model's parameters                                                                                                                                                            |            |               |            |
| # 1 -> direct synthesis parameters                                                                                                                                                                           |            |               |            |
| # 2 -> timing parameters                                                                                                                                                                                     |            |               |            |
| 9) par                                                                                                                                                                                                       | 1          | 2             | 2          |
| # Extent of amplitude perturbation. Its maximum value is the percentage of the<br># voicing gain specified by this parameter. This parameter when non-zero introduces<br># "shimmer" in the speech waveform. |            |               |            |
| 10) shm                                                                                                                                                                                                      | 0.0        | 6.0           | 10.0       |
| # Minimum value of the extent of amplitude perturbation specified as the fraction<br># of the voicing gain                                                                                                   |            |               |            |
| 11) shm_ratio                                                                                                                                                                                                | 0.0        | 1000.0        | 100000.0   |
| # Rate of amplitude perturbation. This filter coefficient of a FOS which is<br># used to filter the amplitude perturbation sequence.                                                                         |            |               |            |
| 12) sflt                                                                                                                                                                                                     | -1.0       | 0.0           | 0.99       |



Table B-IV ... Continued

| # Parameter Name                                                              | Min. Value | Typical Value | Max. Value |
|-------------------------------------------------------------------------------|------------|---------------|------------|
| # Choice of shimmer perturbation measure to be kept constant                  |            |               |            |
| # 1 -> Mean Shimmer (MS)                                                      |            |               |            |
| # 2 -> Shimmer Factor (SF)                                                    |            |               |            |
| # 3 -> Amplitude Shimmer Quotient (APQ)                                       |            |               |            |
| 13) smeas_typ                                                                 | 1          | 2             | 3          |
| # Parameter to add a constant value to the voicing gain                       |            |               |            |
| 14) avadd                                                                     | 0.0        | 0.0           | 100.0      |
| # Voicing gain (in dB)                                                        |            |               |            |
| 15) av                                                                        | 0          | 60            | 100        |
| # Aspiration noise gain (in dB)                                               |            |               |            |
| 16) ah                                                                        | 0          | 60            | 100        |
| # Parameters for amplitude-modulation of noise source                         |            |               |            |
| # Duration of the first part of the amplitude-modulation waveform             |            |               |            |
| # specified as a fraction of pitch-period.                                    |            |               |            |
| 17) offset                                                                    | 0.0        | 0.5           | 1.0        |
| # Duration of the second part of the amplitude-modulation waveform            |            |               |            |
| # specified as a fraction of pitch-period.                                    |            |               |            |
| 18) dur                                                                       | 0.0        | 0.5           | 1.0        |
| # Amplitude of the first part of the amplitude-modulation waveform            |            |               |            |
| 19) amp1                                                                      | 0.0        | 0.5           | 1.0        |
| # Amplitude of the second part of the amplitude-modulation waveform           |            |               |            |
| 20) amp2                                                                      | 0.0        | 0.5           | 1.0        |
| # Extent of pitch perturbation. Its maximum value is the percentage of the    |            |               |            |
| # fundamental frequency specified by this parameter. This parameter when      |            |               |            |
| # non-zero introduces "jitter" in the speech waveform.                        |            |               |            |
| 21) jit                                                                       | 0.0        | 6.0           | 10.0       |
| # Minimum value of the extent of pitch perturbation specified as the fraction |            |               |            |
| # of fundamental frequency                                                    |            |               |            |
| 22) jit_ratio                                                                 | 0.0        | 1000.0        | 100000.0   |
| # Rate of pitch perturbation. This filter coefficient of a FOS which is       |            |               |            |
| # used to filter the pitch perturbation sequence.                             |            |               |            |
| 23) jfilt                                                                     | -1.0       | 0.0           | 0.99       |
| # Choice of jitter perturbation measure to be kept constant                   |            |               |            |
| # 1 -> Mean Jitter (MJ)                                                       |            |               |            |
| # 2 -> Jitter Factor (JF)                                                     |            |               |            |
| # 3 -> Pitch Perturbation Quotient (PPQ)                                      |            |               |            |
| 24) jmeas_typ                                                                 | 1          | 2             | 3          |

Table B-IV ... Continued

| # Parameter Name                                                                        | Min. Value | Typical Value | Max. Value |
|-----------------------------------------------------------------------------------------|------------|---------------|------------|
| # Parameter to add a constant value to the fundamental frequency                        |            |               |            |
| 25) f0add                                                                               | 0.0        | 0.0           | 100.0      |
| # Scale factor to multiply the fundamental frequency                                    |            |               |            |
| 26) f0scl                                                                               | 0.0        | 1.0           | 100.0      |
| #                                                                                       |            |               |            |
| # Choices for interpreting the values of the gain parameter "av"                        |            |               |            |
| # 1 -> total power in the glottal source pulse                                          |            |               |            |
| # 2 -> total energy in the glottal source pulse                                         |            |               |            |
| # 3 -> peak amplitude of the glottal source pulse                                       |            |               |            |
| # 4 -> peak negative amplitude of the differentiated glottal source pulse               |            |               |            |
| # 5 -> cannot be used for this model                                                    |            |               |            |
| # 6 -> these parameters have no effect on the glottal source pulse                      |            |               |            |
| 27) typ_gain                                                                            | 1          | 1             | 6          |
| # Scale factor for controlling the amplitude of the glottal source pulse without        |            |               |            |
| # changing the values of the "av" parameter                                             |            |               |            |
| 28) scale                                                                               | 0.0        | 1.0           | 1000.0     |
| # Flag to indicate the simulation of perceptual effect of varying "stress" with varying |            |               |            |
| # "pitch." When set, the glottal source pulse is scaled by the value of fundamental     |            |               |            |
| # frequency parameter.                                                                  |            |               |            |
| 29) F0                                                                                  | 0          | 1             | 1          |
| #                                                                                       |            |               |            |

Table B-V

Table of parameters for the Single Pulse Model,  
Multiple Pulse Model and the Impulse Train Model  
with minimum, typical and maximum values

| # Parameter Name                                                                        | Min. Value | Typical Value | Max. Value |
|-----------------------------------------------------------------------------------------|------------|---------------|------------|
| # Parameters for amplitude-modulation of noise source                                   |            |               |            |
| # Duration of the first part of the amplitude-modulation waveform                       |            |               |            |
| # specified as a fraction of pitch-period.                                              |            |               |            |
| 1) offset                                                                               | 0.0        | 0.5           | 1.0        |
| # Duration of the second part of the amplitude-modulation waveform                      |            |               |            |
| # specified as a fraction of pitch-period.                                              |            |               |            |
| 2) dur                                                                                  | 0.0        | 0.5           | 1.0        |
| # Amplitude of the first part of the amplitude-modulation waveform                      |            |               |            |
| 3) amp1                                                                                 | 0.0        | 0.5           | 1.0        |
| # Amplitude of the second part of the amplitude-modulation waveform                     |            |               |            |
| 4) amp2                                                                                 | 0.0        | 0.5           | 1.0        |
| # Choices for interpreting the values of the gain parameter "av"                        |            |               |            |
| # 1 -> total power in the glottal source pulse                                          |            |               |            |
| # 2 -> total energy in the glottal source pulse                                         |            |               |            |
| # 3 -> peak amplitude of the glottal source pulse                                       |            |               |            |
| # 4 -> peak negative amplitude of the differentiated glottal source pulse               |            |               |            |
| # 5 -> cannot be used for this model                                                    |            |               |            |
| # 6 -> these parameters have no effect on the glottal source pulse                      |            |               |            |
| 5) typ_gain                                                                             | 1          | 1             | 6          |
| # Scale factor for controlling the amplitude of the glottal source pulse without        |            |               |            |
| # changing the values of the "av" parameter                                             |            |               |            |
| 6) scale                                                                                | 0.0        | 1.0           | 1000.0     |
| # Flag to indicate the simulation of perceptual effect of varying "stress" with varying |            |               |            |
| # "pitch." When set, the glottal source pulse is scaled by the value of fundamental     |            |               |            |
| # frequency parameter.                                                                  |            |               |            |
| 7) F0                                                                                   | 0          | 1             | 1          |
| # Name of the sampled data file with waveform of glottal source pulse(s)                |            |               |            |
| 8) fil_nam*                                                                             | xxxxxx     | gw1_def.d     | xxxxxx     |
| #                                                                                       |            |               |            |

\* This parameter is not specified when the impulse train is used as a glottal source model.

# There are no minimum and maximum sampled data files.

## APPENDIX C SOFTWARE AND FLOW CHART

### Software/Hardware Description

Software for the flexible formant synthesizer is implemented in the "C" programming language. To add flexibility to our software, the "C" programming language was found to be more suitable than FORTRAN77 programming language. The "C" programming language is very useful for writing compact and efficient programs. Readers are referred to the book "The C programming language" by Brian W. Kernighan and Dennis M. Ritchie (1978) for details on various features of the "C" programming language. Among several useful features of "C" programming language exploited in our software, we have highlighted a few by giving examples.

The synthesizer architecture is modular (can be split into sections) and each module has a specific structure. By using the "C" programming language it is easy to implement each module as a single, independent object and maintain its structure by creating a software structure with appropriate fields. For example, a resonator (structure) along with its properties (fields) is treated as one object in the software. Each second order resonator is implemented by a software structure with one field for resonator input, one for resonator output, three fields for storing the three resonator coefficients. Each structure or an array of data or structures can be accessed by its pointer, i.e., by the address of its memory location. Both the features, structures and pointers, were useful for implementing the flexible filter banks in the flexible formant synthesizer. For example, the structures for resonators and anti-resonators,

array of pointers these resonator and anti-resonator structures are used to create a filter bank with a variable number of resonators and anti-resonators. The rearrangement of resonators and anti-resonators in a filter bank can be achieved by adding, removing or rearranging their pointers in the array of pointers for the filter bank.

Adding flexibility to the formant synthesizer requires both a flexible synthesis algorithm and a flexible synthesizer architecture. To achieve this, it is required that the user be provided with multiple choices for selecting a module at each synthesis stage. For example, the user may select a first order FIR (Finite Impulse Response) filter (module) or a first order IIR (Infinite Impulse Response) filter (module) in series with the cascade filter bank to modify the magnitude frequency response of the cascade filter bank. In our software, we implemented each module of the synthesizer architecture by a structure and coded the operational algorithm of that module in a "function." A useful feature of the "C" programming language is that a function can be passed to another function through an argument list of the latter function. We could implement a compact and efficient software by using this feature at all the synthesis algorithm stages, where only one of the several available modules (functions) will be used by the user. For example, the function for the FOS (First Order System) simulated a FIR filter if the function for FIR filter was passed to it or simulated an IIR filter if the function for the IIR filter was passed to it.

On many occasions during the synthesis, the length of data may exceed the length of arrays in which data is supposed to be stored. Such memory shortage problems create "run-time" errors during the synthesis of an utterance. Another useful feature of the "C" programming language is that memory required for arrays and structures can be dynamically allocated during "run-time." In our software, memory required for storing input parameters and for storing data (input, output and intermediate signal) is dynamically allocated during the "run-time" and thus memory shortage

problems are avoided. Also, the input and output statements in the "C" programming language have very few restriction in terms of format of data. Our experience is that the input/output errors due to the data formatting occur less frequently during "run-time," when using the "C" programming language as compared to other programming languages, such as FORTRAN77.

We have implemented the flexible formant synthesizer on a SUN Workstation. This software may work on other mainframes with the "C" programming language compiler but may require some changes in the input/output statements. The libraries for input/output functions are different for different makes of computers. In our software the input and output files may be either in the ESPS ( Entropic Signal Processing System) software format or ASCII format. ESPS software is a speech and general purpose signal processing software package with several utilities for data files management and record keeping. It is marketed by Entropic Speech, Inc. For listening to synthesized sampled data speech, the "splay" program provided in the ESPS software can be used. This program converts the sample data speech to an analog speech signal using the D/A (digital to analog) converter on the SPARC stations (manufactured by SUN microsystems). Unfortunately, this D/A converter has a fixed sampling frequency of 8KHz and has a resolution of only 8 bits. ESPS formatted files can be listened to by using the "wplay" program in the WAVES+ software package (also marketed by Entropic Speech, Inc) on the SUN-3 Workstations. This program converts the sampled-data speech signal to an analogue speech signal using the 16 bit D/A converter on the AT&T's DSP32 board. The user can select the sampling frequency of the sampled data. ASCII formatted files can also be created by the flexible formant synthesizer. For using any other D/A conversion board, the ASCII formatted files can be easily modified as per the requirements.

The input (synthesizer parameters) and output (sampled-data speech) to and from the flexible formant synthesizer software is by means of data files. Various types

of data files are used by the flexible formant synthesizer software to specify the synthesizer parameters to the synthesizer and to store the output synthesized speech. The "Formant Synthesizer User's manual" explains in detail various types of input and output files associated with the flexible formant synthesizer software. This manual is available at Mind-Machine Interaction Research Center at the University of Florida.

#### Flowchart for Formant Synthesizer

Since the flexible formant synthesizer software includes flexibility in parameter specification, synthesis algorithm and synthesizer architecture, it is much bigger than the software for the Klatt's cascade/parallel formant synthesizer. The complete software is coded in approximately 15,000 lines. The input portion of the synthesizer software alone is coded in about 4,000 lines. For such software, it is not possible to create a detailed flowchart, it would be too long and complicated. In Figure C-1 we have shown a simple flowchart for the flexible formant synthesizer software. This flowchart briefly describes the input/output to the flexible formant synthesizer and the synthesis algorithm.

## FORMANT SYNTHESIZER FLOWCHART

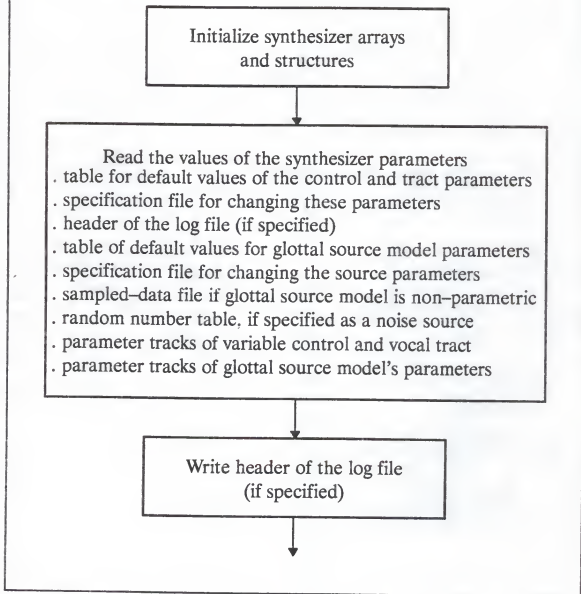


Figure C-1 : Flowchart for flexible formant synthesizer algorithm



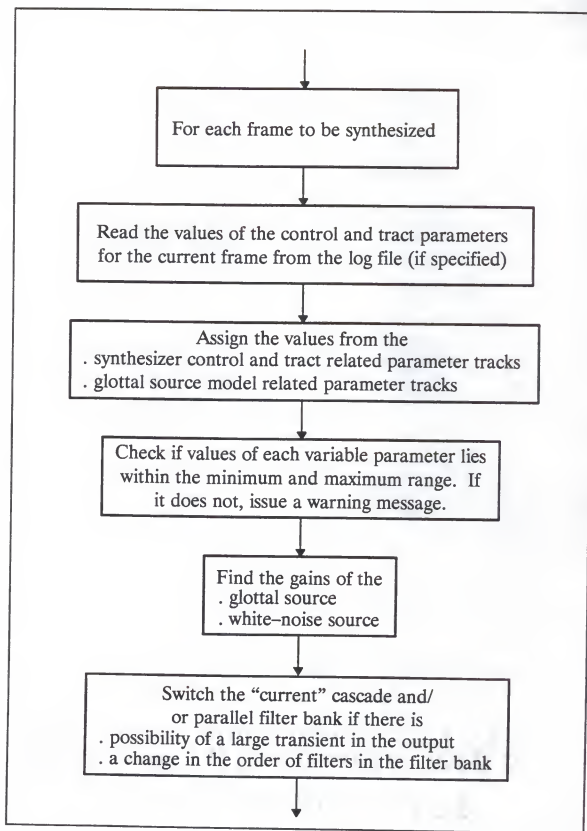


Figure C-1 : continued

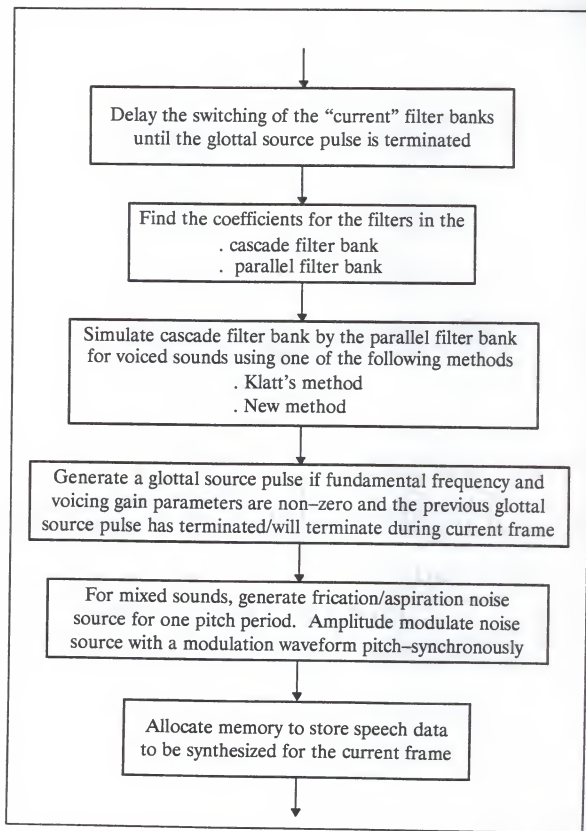


Figure C-1 : continued

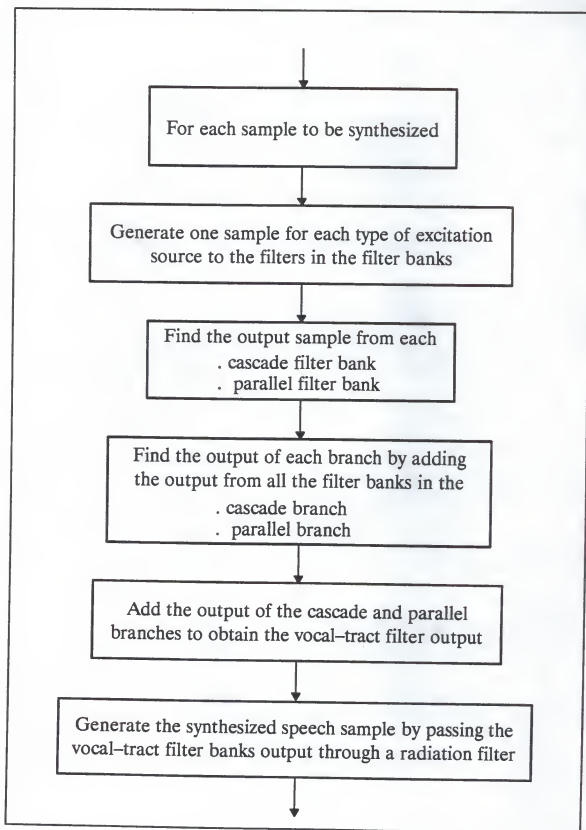


Figure C-1 : continued

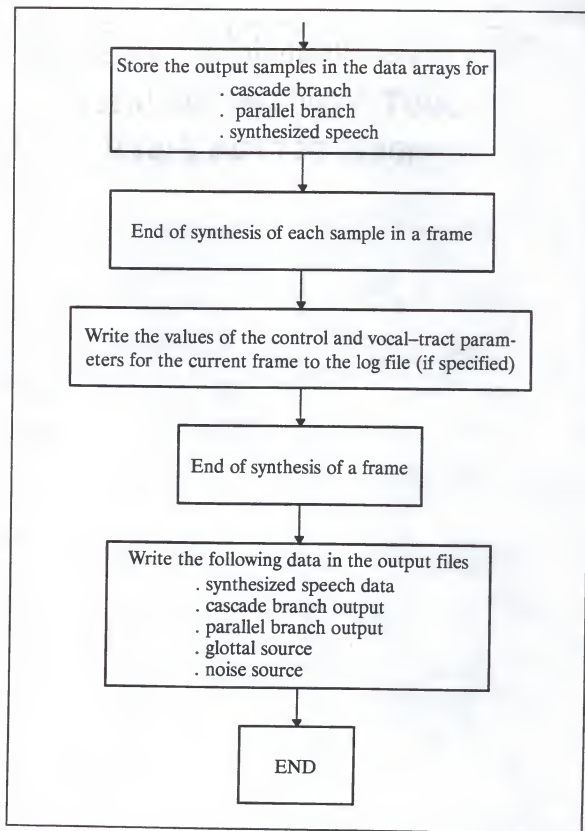


Figure C-1 : continued

## APPENDIX D CONTROLLING JITTER IN SYNTHETIC SPEECH

The Jitter Factor (JF) measure can be defined as the mean of the absolute values of the fluctuations in the fundamental frequency parameter, normalized to the mean value of the fundamental frequency parameter, i.e.,

$$JF = \frac{\frac{1}{N-1} \sum_{i=2}^N |f0_i - f0_{i-1}|}{\frac{1}{N} \sum_{i=1}^N f0_i} \quad D-1$$

where “f0” is the fundamental frequency parameter. A fluctuation in the fundamental frequency parameter is defined as the first order difference in the two consecutive values of the “f0” parameter i.e.,  $f0_i - f0_{i-1}$ .

The Frequency Perturbation Quotient (FPQ) measure can be defined as one third the mean of the absolute value of the rate of change of fluctuations in the fundamental frequency parameter, normalized to the mean value of the fundamental frequency parameter.

$$FPQ = \frac{\frac{1}{N-2} \sum_{i=2}^{N-1} \frac{|(f0_{i+1} - f0_i) - (f0_i - f0_{i-1})|}{3}}{\frac{1}{N} \sum_{i=1}^N f0_i} = \frac{\frac{1}{N-2} \sum_{i=2}^{N-1} \frac{|f0_{i+1} - 2f0_i + f0_{i-1}|}{3}}{\frac{1}{N} \sum_{i=1}^N f0_i} \quad D-2$$

The Directional Jitter (DJ) measure can be defined as the zero crossing rate of the fluctuations in the fundamental frequency parameter.

$$DJ = \frac{(\# \text{ of times the fluctuations in the } f0 \text{ contour change sign})}{\text{Total \# of fluctuations in the } f0 \text{ contour} - 1} \quad D-3$$

The JF, FPQ and DJ measures are related to the perturbation of the fundamental frequency parameter, which, we address as “pitch perturbation” in this study.

When analyzing speech data collected from human subjects, a pitch perturbation sequence is obtained by subtracting the mean value of the fundamental frequency parameter from the fundamental frequency contour. The mean value of the fundamental frequency parameter,  $f0_{mean}$ , is given by

$$f0_{mean} = \frac{1}{N} \sum_{i=1}^N f0_i \quad D-4$$

where  $f0_i$  are the values of the “ $f0$ ” parameter in the fundamental frequency contour. The instantaneous value of pitch perturbation is obtained as

$$per_i = f0_i - f0_{mean} \quad D-5$$

In the new glottal source model, the pitch perturbation sequence is generated by the pitch perturbation source. The pitch perturbation sequence is added to the parameter “ $f0$ ” with constant value to obtain the fundamental frequency contour. The value of the parameter “ $f0$ ” for generating the  $i$ th glottal source pulse is given by

$$f0_i = per_i + f0_{mean} \quad D-6$$

From equations D-5 and D-6, we observe that

$$f0_i - f0_{i-1} = per_i - per_{i-1} \quad D-7$$

$$f0_{i+1} - 2f0_i + f0_{i-1} = per_{i+1} - 2per_i + per_{i-1} \quad D-8$$

The equations D-7 and D-8 show that both the fluctuations and the rate of change of fluctuations in the pitch perturbation sequence are equivalent to the fluctuations and the rate of change of fluctuations in the fundamental frequency parameter, respectively. Thus, the JF and FPQ measures are related to both the extent (range of values) of pitch perturbation, and the rate of change of pitch perturbation in a pitch perturbation sequence. The DJ measure is related to the zero crossing rate (ZCR) of a pitch perturbation sequence, i.e., to the spectrum of the pitch perturbation

sequence. Since, in the new glottal source model, the "extent of pitch perturbation" and the "(zero crossing) rate of pitch perturbation" can be controlled, it is possible to synthesize speech tokens with desired values of JF, FPQ and DJ measures. Using this feature, we can evaluate the effects of varying the JF, FPQ and DJ measures on various vocal characteristics.

In the experiments to evaluate the effect of variations of the JF, FPQ or DJ measures on the vocal characteristics, it is desirable to keep only one of the three JF, FPQ or DJ measures variable while keeping the other two measures constant. From the definitions of the JF and FPQ measures, it can be observed that these two measures are correlated and cannot be varied independently. Both the JF and FPQ measures vary simultaneously whenever the "extent of pitch perturbation" and "rate of pitch perturbation" are varied. The JF and FPQ measures can be varied only by changing the extent of pitch perturbation only, if the spectrum of the pitch perturbation sequence is not changed. The "extent of pitch perturbation" in a pitch perturbation sequence can be changed by multiplying the pitch perturbation sequence by a scale factor. The DJ measure can be varied by changing the "rate of pitch perturbation" of a pitch perturbation sequence, i.e., by changing the spectrum of a pitch perturbation sequence using a lowpass or a highpass filter. However, filtering the pitch perturbation sequence may result in the change in the "extent of pitch perturbation" (due to the gain of the filter), and therefore, in the values of JF and FPQ measures. In order to be able to vary the DJ measure without changing the "extent of pitch perturbation," and the JF and FPQ measures, the pitch perturbation sequence has to be multiplied by a scale factor, which is multiplicative inverse of the fractional change in the "extent of pitch perturbation" caused by the filtering.

In our glottal source model we have incorporated features to change the JF and FPQ measures independent of DJ measure, and conversely, the DJ measure independently of the JF and FPQ measures. Both the JF and FPQ measures can be

controlled by a single parameter, " $f0_{ext}$ ," which specifies the desired "extent of pitch perturbation" as a fraction of the mean value of the fundamental frequency parameter. The higher the value of the " $f0_{ext}$ " parameter, the higher are the values of the JF and FPQ measures. The DJ measures can be controlled by a single parameter, " $a_0$ ," which specifies the value of the coefficient of a first order FIR or a first order IIR filter used for changing the "rate of pitch perturbation" of the pitch perturbation sequence. It is observed that if the spectrum of the pitch perturbation sequence has white-noise characteristics, the ZCR (the zero crossing rate of the pitch perturbation sequence) is approximately equal to 0.5. If the spectrum shows higher magnitude at higher frequencies than at lower frequencies, the ZCR is higher than 0.5. If the spectrum shows higher magnitude at lower frequencies than at higher frequencies, the ZCR is lower than 0.5. The filter type and the bandwidth of the filter determines the "rate of pitch perturbation" of the pitch perturbation sequence and thus, the value of the DJ measure. The DJ measure is defined as the average zero crossing rate of differentiated pitch perturbation sequence, and therefore is related to ZCR. The relationship between the ZCR and the DJ measure of a pitch perturbation sequence is described later. When the pitch perturbation sequence is filtered it is also multiplied by a scale factor to maintain the desired value of "extent of pitch perturbation" and thereby the values of JF and FPQ measures.

### Controlling the JF and FPQ Measures

Pinto and Titze (1990) have unified the various perturbation measures commonly used by several researchers to measure the period-to-period variations in the fundamental frequency. They have unified the JF and FPQ measures in terms of mean rectified (MR) values of the first and second order differentiated pitch perturbation sequences, respectively.



A zeroth order MR value of a pitch perturbation sequence,  $MR^0(per)$ , is defined as

$$MR^0(per) = \frac{1}{N} \sum_{i=1}^N |per_i| \quad D-9$$

The first order MR value of a perturbation sequence,  $MR^1(per)$ , is defined as

$$MR^1(per) = \frac{1}{N-1} \sum_{i=2}^N |per_i - per_{i-1}| \quad D-10$$

The second order MR value of a perturbation sequence,  $MR^2(per)$ , is defined as

$$\begin{aligned} MR^2(per) &= \frac{1}{N-2} \sum_{i=2}^{N-1} |(per_{i+1} - per_i) - (per_i - per_{i-1})| \\ &= \sum_{i=2}^{N-1} |per_{i+1} - 2per_i + per_{i-1}| \end{aligned} \quad D-11$$

Comparing equations D-1, D-2, D-7 and D-8 with D-10 and D-11, we observed that the JF and FPQ measures can be defined in terms of  $MR^1(per)$  and  $MR^2(per)$ . Using the equation D-10, the JF measure can be defined as

$$JF = \frac{MR^1(per) * 100}{f0_{mean}} \quad D-12$$

and using the equation D-11, the FPQ measure can be defined as

$$FPQ = \frac{MR^2(per)}{3 * f0_{mean}} \quad D-13$$

Pinto and Titze (1990) have also shown that for a pitch perturbation sequence having a Gaussian Distribution with zero mean value and standard deviation equal to  $\sigma_{per}$ , the zeroth order MR value ( $MR^0$ ) is given by the equation

$$MR^0(per) = \int_{-\infty}^{\infty} \frac{|per|}{\sqrt{2\pi\sigma_{per}^2}} e^{-\frac{(per)^2}{2\sigma_{per}^2}} dper \quad D-14$$

This equation can be simplified to

$$MR^0(per) \approx 0.8\sigma_{per} \quad D-15$$

The  $MR^1(per)$  and the  $MR^2(per)$  measures are the mean rectified values of the perturbation sequences obtained by taking the first and second order difference of the pitch perturbation sequences, i.e., by filtering the pitch perturbation sequence by a first order FIR and by a cascade of two first order FIR filters (with filter coefficient equal to  $-1$ ), respectively. The relationship between the  $MR^0(per)$ ,  $MR^1(per)$  and  $MR^2(per)$  values of a pitch perturbation sequence is shown in the Figure D-1a. If the value of  $MR^0(per)$  of a pitch perturbation sequence is known, we can find the  $MR^1(per)$  and  $MR^2(per)$  values and also the values of the JF and FPQ measures analytically without actually carrying out the filtering of the pitch perturbation sequence. The analytical relationships between the  $MR^0(per)$ ,  $MR^1(per)$  and  $MR^2(per)$  values of a pitch perturbation sequence can be found by taking in to account the following facts:

- 1) If a pitch perturbation sequence at the input of a linear filter has a pseudo Gaussian distribution with zero mean, the output perturbation sequence also has a pseudo Gaussian distribution with zero mean. The standard deviation of the output sequence may be different from the standard deviations of the input pitch perturbation sequence due to the gain of the filter.
- 2) If the pitch perturbation sequence has white-noise spectral characteristics, the power in the output perturbation sequence is  $\sum |h(n)|^2$  times the power in the input pitch perturbation, where  $h(n)$  is the impulse response of the filter bank. If the samples of the pitch perturbation sequence are uncorrelated and have zero mean value, the standard deviation of the perturbation sequence at the output of a filter bank is square root of  $(\sum |h(n)|^2)$  times the standard deviation of the pitch perturbation sequence at the input.

Therefore, if the standard deviation of the pitch perturbation sequence is  $\sigma_{per}$  and the impulse response of a first order differentiator is represented as

$$h(n) = \delta(n) - \delta(n-1) \quad D-16$$

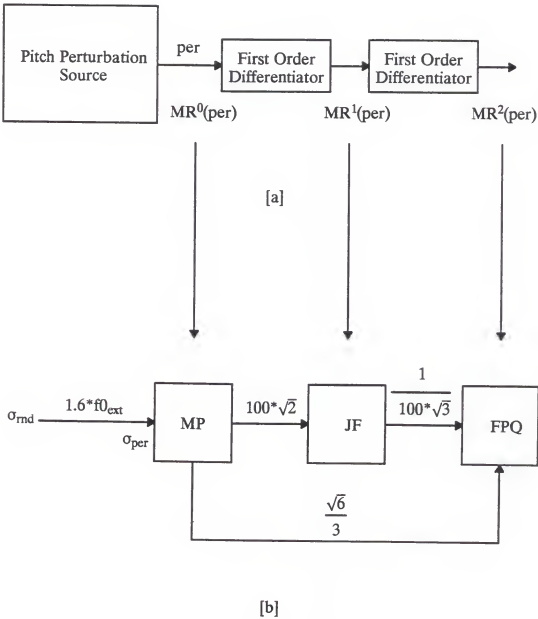


Figure D-1: a) Relationship between pitch perturbation source and  $MR^0$ ,  $MR^1$  and  $MR^2$  values  
 b) Relationship between  $\sigma_{rnd}$  and MP, JF and FPQ measures

the standard deviation of the output perturbation sequence is  $\sqrt{2} * \sigma_{per}$ . The impulse response of a cascade of two first order differentiator is represented as

$$h(n) = \delta(n) - 2\delta(n-1) + \delta(n-2) \quad D-17$$

Therefore, the standard deviation of the output perturbation sequence is  $\sqrt{6} * \sigma_{per}$ . The approximate values of  $MR^1(per)$  and  $MR^2(per)$  can be derived using the equation D-15 as follows

$$MR^1(per) = \sqrt{2} * MR^0(per) \approx 0.8 * \sqrt{2} * \sigma_{per}$$

D-18

and

$$MR^2(per) = \sqrt{6} * MR^0(per) \approx 0.8 * \sqrt{6} * \sigma_{per}$$

D-19

Substituting these approximations in equations D-12 and D-13, the JF measure can be expressed in terms of  $\sigma_{per}$  as

$$JF = \frac{MR^1(per) * 100}{f0_{mean}} \approx \frac{100 * 0.8 * \sqrt{2} * \sigma_{per}}{f0_{mean}} \quad D-20$$

and the FPQ measure can be expressed in terms of  $\sigma_{per}$  as

$$FPQ = \frac{MR^2(per)}{3 * f0_{mean}} \approx \frac{0.8 * \sqrt{6} * \sigma_{per}}{3 * f0_{mean}} \quad D-21$$

For describing the JF and FPQ measures, we define a new measure "mean rectified perturbation," MP, such that

$$MP = \frac{MR^0(per)}{f0_{mean}} \approx \frac{0.8 * \sigma_{per}}{f0_{mean}} \quad D-22$$

Using the equations D-20, D-21 and D-22, the the JF and FPQ measures can be given in terms of the MP measure as

$$JF = \sqrt{2} * 100 * MP \quad D-23$$

and

$$FPQ = \frac{\sqrt{6} * MP}{3} \quad D-24$$

From the equations D-22, D-23 and D-24, we can observe that the MP, JF and FPQ measures can be controlled by the standard deviation of the pitch perturbation sequence,  $\sigma_{per}$ .

In the new glottal source model, the standard deviation of the pitch perturbation sequence,  $\sigma_{per}$ , can be controlled by the "extent of pitch perturbation" specified by the parameter " $f0_{ext}$ ." The random number sequence generated by the random number generator in the new glottal source model has a pseudo Gaussian distribution within the range  $\pm 0.5$  and white-noise spectral characteristics. All of the random number sequences (with different seed values) generated by the random number generator have approximately the same standard deviation. Let  $\sigma_{rnd}$  represent the standard deviation of the random number sequence. (The value of the standard deviation,  $\sigma_{rnd}$ , is specific to the random number generator being used.) The random number sequence is multiplied by a scale factor,  $2 * f0_{ext} * f0_{mean}$ , in order to obtain the pitch perturbation sequence that has a pseudo Gaussian distribution within the range  $\pm f0_{ext} * f0_{mean}$  and a zero mean value. The standard deviation,  $\sigma_{per}$ , of the pitch perturbation sequence generated by the new glottal source model is equal to

$$\sigma_{per} = 2 * f0_{ext} * f0_{mean} * \sigma_{rnd} \quad D-25$$

Substituting the value of the standard deviation,  $\sigma_{per}$ , in equations D-22, D-23 and D-24, we obtain

$$MP = \frac{MR^0(per)}{f0_{mean}} = \frac{0.8 * \sigma_{per}}{f0_{mean}} = 1.6 * f0_{ext} * \sigma_{rnd} \quad D-26$$

$$JF = \frac{MR^1(per) * 100}{f0_{mean}} = \frac{100 * 0.8 * \sqrt{2} * \sigma_{per}}{f0_{mean}} \quad D-27$$

$$= 100 * \sqrt{2} * 1.6 * f0_{ext} * \sigma_{rnd}$$

and

$$FPQ = \frac{MR^2(per)}{3 * f0_{mean}} = \frac{0.8 * \sqrt{6} * \sigma_{per}}{3 * f0_{mean}} = \frac{\sqrt{6} * 1.6 * f0_{ext} * \sigma_{rnd}}{3} \quad D-28$$

We can observe that the MP, JF and FPQ measures are independent of  $f0_{\text{mean}}$ , since the pitch perturbation sequence is scaled by  $f0_{\text{mean}}$ . Since the value of  $\sigma_{\text{rnd}}$  is constant, the values of the MP, JF and FPQ measures depend only upon the “extent of pitch perturbation” parameter, “ $f0_{\text{ext}}$ .” Thus using the new glottal source model, it is possible to synthesize a speech signal with desired values of MP, JF and FPQ measures. The values of these three measures for the given value of “ $f0_{\text{ext}}$ ” parameter are given by the above equations. The relationships between MP, JF and FPQ measures and  $\sigma_{\text{rnd}}$  are described in Figure D-1b. The analytical expressions for the MP, JF and FPQ measures are listed in Table D-I.

### Controlling the DJ Measure

In order to vary the DJ measure, the spectrum of the pitch perturbation sequence should be altered by filtering the pitch perturbation sequence. The type of spectral-shaping filter (lowpass or highpass) determines the change in the shape of the original spectrum of the pitch perturbation sequence. The filter coefficient, “ $a_{\text{fl}}$ ,” determines the range of frequencies in which the spectrum of pitch perturbation sequence may have higher magnitude. By varying the value of the parameter “ $a_{\text{fl}}$ ” we can vary the ZCR of the pitch perturbation sequence, and thereby, vary the DJ measure of the pitch perturbation sequence. It should be noted that the ZCR is the zero crossing rate of the pitch perturbation sequence and the DJ measure is the zero crossing rate of the differentiated pitch perturbation sequence.

For finding the relationship between the filter coefficient, “ $a_{\text{fl}}$ ,” ZCR and DJ measure, we generated 10 pitch perturbation sequences, each with 2000 samples, and filtered each of the 10 pitch perturbation sequences. The filter coefficient, “ $a_{\text{fl}}$ ,” was varied from -1.0 to 1.0 in steps of 0.2. For the range of values of the parameter, “ $a_{\text{fl}}$ ,” from -1.0 to 0.0 (i.e.,  $-1.0 \leq a_{\text{fl}} < 0.0$ ), the pitch perturbation sequence was filtered by a highpass filter. When  $a_{\text{fl}} = 0.0$ , the pitch perturbation sequence was not filtered.

Table D-I

The analytical expressions for the MP, JF and FPQ measures

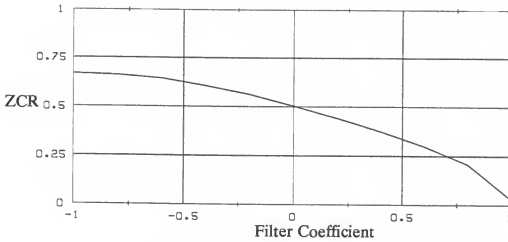
| MP                                               | JF                           | FPQ                              |
|--------------------------------------------------|------------------------------|----------------------------------|
| $1.6 * f_{0_{\text{ext}}} * \sigma_{\text{rnd}}$ | $100 * \sqrt{2} * \text{MP}$ | $\frac{\sqrt{6} * \text{MP}}{3}$ |

For the range of values of the parameter, " $a_{\text{ff}}$ ," from 0.0 to 1.0 (i.e.,  $0.0 < a_{\text{ff}} < 1.0$ ), the pitch perturbation sequence was filtered by a lowpass filter. The mean value and standard deviation of the ZCR and DJ measure of the 10 pitch perturbation sequences were calculated for each value of the " $a_{\text{ff}}$ " parameter. Figure D-2 shows the variation of the ZCR and DJ measure with the variation of the value of the filter coefficient from -1.0 to 1.0 in steps of 0.2. We can observe that the value of ZCR is above 0.5 when the pitch perturbation sequence is filtered by a highpass filter, approximately equal to 0.5 when the pitch perturbation sequence is not filtered and less than 0.5 when the pitch perturbation sequence is filtered by a lowpass filter. The DJ measure varies proportionately with the ZCR, although the range of variation is not as large. It is observed that the DJ measure increases as the bandwidth of the highpass filter decreases. The DJ measure decreases as the bandwidth of the lowpass filter decreases.

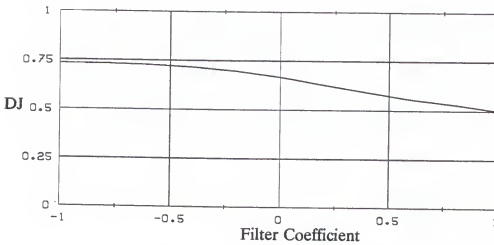
#### Effect of Varying the DJ Measure on the JF and FPQ Measures

Due to the addition of a spectral-shaping filter to the pitch perturbation source, there is a change in the relationship between the  $\text{MR}^0$ ,  $\text{MR}^1$  and  $\text{MR}^2$  values and the standard deviation of the pitch perturbation sequence,  $\sigma_{\text{per}}$ . The block diagram in Figure D-3 shows the relationship between the modified pitch perturbation source and the  $\text{MR}^0$ ,  $\text{MR}^1$  and  $\text{MR}^2$  values. As seen earlier, for the input with white noise characteristics the standard deviation of the output (filtered) perturbation sequence is equal to the square root of  $(\sum |h(n)|^2)$  times the standard deviation of the pitch perturbation sequence at the input, where  $h(n)$  is the impulse response of the filter bank. Thus, the  $\text{MR}^0$ ,  $\text{MR}^1$  and  $\text{MR}^2$  values of a pitch perturbation sequence values can be obtained in terms of  $\sigma_{\text{per}}$  by taking into account the impulse response of the spectral shaping filter, the impulse response of the cascade of spectral-shaping filter





← Highpass →      ← Lowpass →  
 Decreasing Bandwidth ←      → Decreasing Bandwidth



← Highpass →      ← Lowpass →  
 Decreasing Bandwidth ←      → Decreasing Bandwidth

Figure D-2: a) Zero Crossing Rate (ZCR) versus filter coefficient  
 b) Directional Jitter (DJ) versus filter coefficient

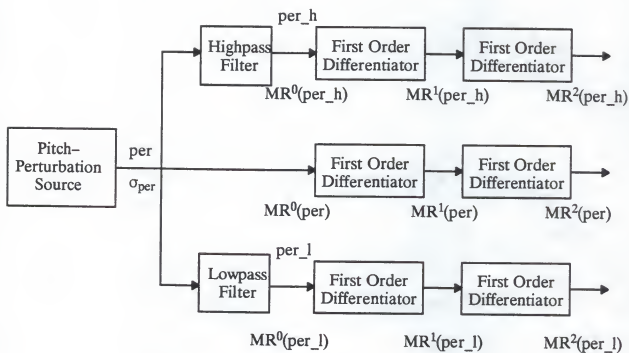


Figure D-3: Relationship between  $MR^0$ ,  $MR^1$  and  $MR^2$  values and the pitch perturbation source when it is filtered

and a single first order differentiator, and the impulse response of the cascade of spectral-shaping filter and two first order differentiators, respectively.

The impulse response of a first order FIR filter with coefficient " $a_0$ " is given by

$$h(n) = \delta(n) - a_0 \delta(n-1) \quad \text{D-29}$$

The impulse response of a first order IIR filter (normalized to 0 dB at dc) is given by

$$h(n) = (1-a_0) * a^n u(n) \quad \text{D-30}$$

Therefore, the standard deviation of a highpass filtered pitch perturbation sequence,  $\sigma_{per\_h}$ , is given by

$$\sigma_{per\_h} = \sqrt{1 + a_0^2} * \sigma_{per} \quad \text{D-31}$$

The standard deviation of a lowpass filtered pitch perturbation sequence,  $\sigma_{per\_l}$ , is given by

$$\sigma_{per\_l} = \sqrt{\frac{(1-a)^3}{1-a^2}} * \sigma_{per} = \sqrt{\frac{(1-a)^2}{1+a}} * \sigma_{per} \quad \text{D-32}$$

Using the equation D-31 and D-32, the zeroth order MR value of a highpass filtered pitch perturbation sequence,  $MR^0(per\_h)$ , and the zeroth order MR value of a lowpass filtered pitch perturbation sequence,  $MR^0(per\_l)$ , are given by

$$MR^0(per\_h) = \sqrt{1 + a_0^2} * MR^0(per) \quad \text{D-33}$$

and

$$MR^0(per\_l) = \sqrt{\frac{1-a_0}{1+a_0}} * MR^0(per) \quad \text{D-34}$$

The first order MR value of a highpass filtered pitch perturbation sequence,  $MR^1(per\_h)$ , is obtained from the impulse response of the cascade of the highpass filter and a first order differentiator. After several steps of algebraic manipulation, we obtain

$$MR^1(per\_h) = \sqrt{(1-a_0 + a_0^2)} * MR^1(per) \quad \text{D-35}$$

Similarly, for the lowpass filtered pitch perturbation sequence,

$$MR^1(per\_l) = \sqrt{\frac{(1-a_0)^2}{1+a_0}} * MR^1(per) \quad D-36$$

The second order MR value of a highpass filtered pitch perturbation sequence,  $MR^2(per\_h)$ , and the second order MR value of a lowpass filtered pitch perturbation sequence,  $MR^2(per\_l)$ , after several steps of algebraic manipulations, are obtained as

$$MR^2(per\_h) = \sqrt{\frac{(3-4a_0+3a_0^2)}{3}} * MR^2(per) \quad D-37$$

and

$$MR^2(per\_l) = \sqrt{\frac{(1-a_0)^2(3-a_0)}{3(1+a_0)}} * MR^2(per) \quad D-38$$

Accordingly, the values of MP, JF and FPQ measures when the pitch perturbation sequence is highpass filtered or lowpass filtered, in terms of their corresponding values when the pitch perturbation sequence is not filtered, i.e., in terms of the values given by equations D-26, D-27 and D-28, are given by

$$MP\_h = \sqrt{(1+a_0^2)} * MP \quad D-39$$

and

$$MP\_l = \sqrt{\frac{1-a_0}{1+a_0}} * MP \quad D-40$$

$$JF\_h = \sqrt{(1-a_0+a_0^2)} * JF \quad D-41$$

and

$$JF\_l = \sqrt{\frac{(1-a_0)^2}{1+a_0}} * JF \quad D-42$$

$$FPQ\_h = \sqrt{\frac{(3-4a_0+3a_0^2)}{3}} * FPQ \quad D-43$$

and

$$FPQ\_l = \sqrt{\frac{(1-a_0)^2(3-a_0)}{3(1+a_0)}} * FPQ \quad D-44$$

Thus, the spectral-shaping filter not only changes the DJ measure of the pitch perturbation sequence but also the values of the MP, JF and FPQ measures. When the pitch perturbation sequence is filtered, the values of the JF and FPQ measures depend upon the "extent of pitch perturbation," the filter type and the filter coefficient, i.e., upon the parameter, " $f_{0\text{ext}}$ ," and the filter coefficient, " $a_{f0}$ ." The scale factors (the terms with the parameter " $a_{f0}$ ") in the above equations indicate the fractional change in the values of the MP, JF and FPQ measures due to filtering of the pitch perturbation sequence. Each scale factor in these equations depend upon the type of filter and the value of the filter coefficient. These scale factors are listed in Table D-II.

With a variation in the value of the filter coefficient and/or the change in the type of filter used, the values of all the MP, JF and FPQ measures will vary. However, we can keep at least one of the MP, JF and FPQ measures constant while the DJ measure is varied. This can be achieved by multiplying the pitch perturbation sequence by the multiplicative inverse of the scale factor, selected from Table D-II. With this method, it is not possible to keep the values of all the MP, JF and FPQ measures constant when the DJ measure is varied. Only one of the MP, JF and FPQ measures can be kept independent of variations in DJ measure. The values of the other two measures will vary with variations in the DJ measure.

For testing the analytical expressions in the Table D-I and the scale factors in the Table D-II, we generated 10 pitch perturbation sequences, each with 2000 samples. The mean value and standard deviation of each of the MP, JF and FPQ measures and the values of the JF and FPQ measures in terms of the MP measure were calculated from the unfiltered pitch perturbation sequences. Also, the values of the JF and FPQ measures in terms of MP measure were calculated using the analytical expressions from the Table D-I. Then the 10 pitch perturbation sequences were filtered by a lowpass filter ( $a_{f0}=0.8$ ) and the mean value and standard deviation of each of the MP, JF and FPQ measures were calculated. Then the 10 pitch perturbation sequences

Table D-II

The fractional changes (scale factors) in the values of the MP, JF and FPQ measures when the pitch perturbation sequence is highpass and lowpass filtered

| Measure<br>Filter                         | MP                                     | JF                                         | FPQ                                                       |
|-------------------------------------------|----------------------------------------|--------------------------------------------|-----------------------------------------------------------|
| Highpass filter<br>$-1.0 \leq a_{f0} < 0$ | $\sqrt{1 + a_{f0}^2}$                  | $\sqrt{(1 - a_{f0} + a_{f0}^2)}$           | $\sqrt{\frac{(3 - 4a_{f0} + 3a_{f0}^2)}{3}}$              |
| Lowpass filter<br>$0 < a_{f0} < 1.0$      | $\sqrt{\frac{1 - a_{f0}}{1 + a_{f0}}}$ | $\sqrt{\frac{(1 - a_{f0})^2}{1 + a_{f0}}}$ | $\sqrt{\frac{(1 - a_{f0})^2(3 - a_{f0})}{3(1 + a_{f0})}}$ |

Table D-III

Values of the JF and FPQ measures from the analytical expression in NO TAG and values of the scale factors in NO TAG

| Measure         | MP         |                  | JF         |                   | FPQ        |                    |
|-----------------|------------|------------------|------------|-------------------|------------|--------------------|
| Filter          | Analytical | Simulated        | Analytical | Simulated         | Analytical | Simulated          |
| No filter       | 1.0        | 1.0              | 1.4142     | 1.4079<br>(1.7)   | 0.8185     | 0.8122<br>(0.0127) |
| Highpass filter | 1.281      | 1.276<br>(0.015) | 2.209      | 2.198<br>(0.034)  | 4.03       | 4.007<br>(0.074)   |
| Lowpass filter  | 1.667      | 1.674<br>(0.056) | 1.054      | 1.054<br>(0.0038) | 1.564      | 1.558<br>(0.018)   |

MP =  $1.6 \cdot f_{0\text{ext}} \cdot \sigma_{\text{rnd}}$

Analytical data MP = 0.116 when  $f_{0\text{ext}} = 1.0$  and  $\sigma_{\text{rnd}} = 0.0723$

Filter coefficient for the lowpass filter is equal to 0.8 and for the highpass filter is equal to -0.8

Values in the parenthesis are the standard deviation

were filtered by a highpass filter ( $a_{f0} = -0.8$ ) and the mean value and standard deviation of each of the MP, JF and FPQ measures were calculated. Then the fractional changes in the mean values of the MP, JF and FPQ measures due to each type of filtering were calculated. Also, the value of each scale factor listed in the Table D-II was calculated by substituting the appropriate value of the filter coefficient. The first row in the Table D-III shows the values of the JF and FPQ measures in terms of the MP measure as calculated from the analytical expressions in the Table D-I and those calculated from the simulated data. The second and the third rows in the same table show the values of the scale factors as calculated from the scale factors in Table D-II and the fractional changes in the values of the MP, JF and FPQ measures due to filtering as calculated from the simulated data. It can be observed from this table that the values obtained analytically (from the analytical expressions and the scale factors from the tables) and those obtained from the simulated data are very close and the standard deviation of each of the mean values calculated from the simulated data is extremely small in each case.

In our glottal source model, when the DJ measure is varied the value of only one of the MP, JF and FPQ measures can be kept constant. The parameter "jmeas\_typ" is used to select which one of the MP, JF and FPQ measures has to be kept constant. If selected, the value of the MP measure is kept constant at  $1.6 \cdot f0_{\text{ext}} \cdot \sigma_{\text{rnd}}$ . If selected, the value of the JF measure is kept constant at  $\sqrt{2} \cdot 1.6 \cdot f0_{\text{ext}} \cdot \sigma_{\text{rnd}}$ . If selected, the value of the FPQ measure is kept constant at  $\sqrt{6} \cdot 1.6 \cdot f0_{\text{ext}} \cdot \sigma_{\text{rnd}}$ . The values of other two pitch perturbation measures will be scaled by the same amount in addition to the fractional change caused by filtering.

## APPENDIX E HARMONIC TO NOISE RATIO

Following the work of Yanagihara (1967), several researchers have proposed methods to measure the degradation of high-frequency harmonics observed in the spectrograms of the speech signal collected from human subjects with breathy and hoarse voices. Of the several possible reasons for degradation of high-frequency harmonics the most commonly cited are: 1) increase in the high-frequency noise, 2) pitch period perturbation (period to period variations of fundamental frequency) and 3) amplitude perturbation (period to period variations of peak amplitude) in the speech signal. The proposed time domain measures are: Harmonic to Noise Ratio (HNR) [Yumoto et al., 1982 and 1984] and the frequency domain measures are: Signal to Noise Ratio (SNR) [Kojima et al., 1980], Relative Harmonic Intensity ( $H_r$ ) [Hiraoka et al., 1984], Noise to Signal Ratio (NSR) [Muta et al., 1988] and Noise to Harmonic Ratio (NHR) [Lee and Childers, 1989]. These and other studies [Eskenazi et al., 1990] have showed differences in the values of these measures for normal and pathological subjects, as well as, correlations with the subjective rating of breathiness and hoarseness severity. Also, some of these studies have correlated the changes in the value of the proposed measure obtained from the pre- and postlaryngeal surgery phonations to the improvement in voice after the surgery.

Recently, Lee and Childers (1989) have defined Noise to Harmonic Ratio (NHR) as the ratio of power in between the harmonics (inter-harmonic components) to the power in the harmonics of the speech signal. In this method, the fundamental frequency,  $F_0$ , of a speech signal segment of a sustained vowel is first estimated from the EGG (electroglottogram) signal [Childers et al., 1990] and then measured using



a frequency domain method. The FFT (Fast Fourier Transform) of the speech signal with a duration of 0.2048 s (2048 samples at 10 KHz) is calculated; a Hamming window is used. The spectrum of the windowed speech signal is divided into regions of width  $F0 \pm F0/2$  along the frequency axis. In each region, the harmonic and inter-harmonic components are defined as shown in Figure E-1. The Noise to Harmonic measure is defined as

$$NHR = 20.0 * \log \left[ \frac{\sum_i N_i}{\sum_i H_i} \right] \quad E-1$$

where  $H_i$  is the harmonic power in the region centered at  $i * F0$  with a bandwidth corresponding to that of the Hamming window and  $N_i$  represents the power in the rest of the region. It was observed that a reliable prediction of breathiness was achieved when the NHR was measured from the harmonics above 2 KHz [Lee, 1988]

Muta et al. (1988) have argued that several methods require a long phonation of a sustained vowel for analysis (approximately 0.2 s), and thus, are sensitive to fluctuations of pitch, intensity, articulation and vibrato. Any of these factors would contribute to an apparent reduction of the harmonic structure of the voice. Also, the reliability of these methods depends on the subjects' ability to produce a long sustained vowel at constant pitch and intensity. They proposed a method to measure the Noise to Signal Ratio (NSR) from the speech signal. The fundamental frequency,  $F0$ , of a speech signal is first estimated using both the time and frequency domain methods. The DFT (Discrete Fourier Transform) of the speech signal with a duration of exactly four pitch periods is calculated; a Hanning window is used. Because the Hanning window covers exactly four pitch periods, harmonic peaks and valleys appear every fourth sample of the DFT. The smallest value of the four DFT samples,  $P(4i-1)$ ,  $P(4i)$ ,  $P(4i+1)$  and  $P(4i+2)$  for the  $i$ th harmonic is assigned as the noise component,  $P_N(i)$ , of each of these four DFT samples. The NSR measure is defined as

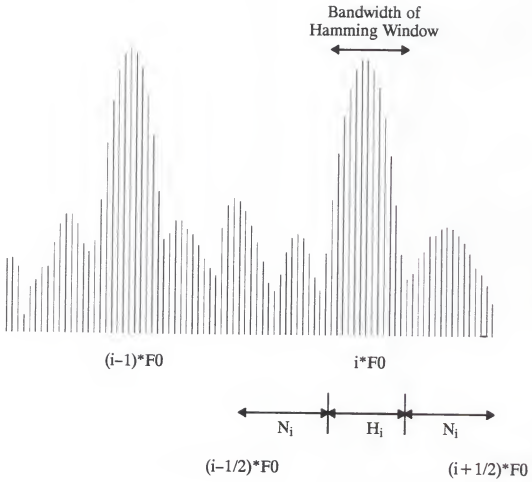


Figure E-1: Harmonic and the inter-harmonic (noise) components

$$NSR = 20.0 * \log \left[ \frac{\sum_{i=3}^{4L+2} P_N(i)}{\sum_{i=3}^{4L+2} P(i)} \right] \quad E-2$$

where  $L$  is the number of harmonics used for calculating the HNR.

They showed that this measure is less sensitive to variations of fundamental frequency than the Harmonic to Noise Ratio defined by Yumoto et al. (1982). Also, the comparison of the pre- and postoperative voices of six patients showed that NSR for the vowel /u/ in continuous speech consistently improved after surgery for all patients, in agreement with their successful therapeutic results.

The method for calculating the NHR [Lee and Childers, 1989] and the NSR [Muta et al., 1988] have both advantages and disadvantages. The advantage of the method proposed by Muta et al., (1988) is that short segments of vowels from continuous speech can be used for analysis. This method is pitch synchronous, sensitive to time-varying changes in the pitch and amplitude perturbation characteristics in continuous speech and does not involve averaging of data from a large number of pitch periods. The disadvantage is that by calculating the DFT of four pitch periods, only four samples of the spectrum of a speech signal segment are available for calculation of the signal and noise power per harmonic. The advantage of the method proposed by Lee and Childers (1989) is that for the same speech signal segment, the FFT samples the spectrum of a speech signal segment more finely than the DFT, and therefore, the power in the harmonic and inter-harmonic components can be measured more accurately than that measured using the DFT. However, this method uses a long window of fixed size, and therefore, is not pitch-synchronous and averages data from many pitch periods.

In this study we propose a pitch-synchronous procedure for the measurement of degradation of high-frequency harmonics. In this procedure we combine both the

pitch-synchronous procedure proposed by Muta et al. (1988) and the NHR proposed by Lee and Childers (1989). We determine the fundamental frequency of a segment of the speech signal using both the time and frequency domain methods. The FFT of a speech signal segment with a specified number of pitch periods is calculated; a Hanning window is used. The advantages of using the FFT are: 1) high speed calculation of HNR and 2) high resolution of speech signal and glottal source waveform, giving an accurate measure of HNR. The HNR is defined as

$$HNR = 20.0 * \log \left[ \frac{\sum_i H_i}{\sum_i N_i} \right] \quad E-3$$

Note that this definition of HNR is different from the HNR measure proposed by Yumoto et al. (1982 and 1984). Also, this definition is inverse of the NHR measure proposed by Lee and Childers (1989). The appropriate values of number of pitch periods of the speech signal for the calculation of the FFT and the starting harmonic for the summation in the equation E-3 for the calculation of HNR are still under further investigation. Muta et al. (1988) have used four pitch periods of the speech signal to measure the NSR. Lee and Childers have used harmonics above 2 KHz to measure the NHR from breathy phonations. Klatt and Klatt, (1990) have reported the presence of aspiration noise in breathy phonations in the high-frequency region starting below 2 KHz. Analysis of several speech signal segments is essential for finding the appropriate values of these parameters.

All the previous studies for measurement of degradation of high-frequency harmonics dealt with the speech signal, since it is easier to find the values of these measures directly from the speech signal than from the glottal flow waveform. Recently, several researchers have described various vocal disorders by their source characteristics. Klatt and Klatt, (1990) have described breathy and creaky vocal characteristics in terms of glottal source characteristics required for synthesis. Gobl

(1989) has described the source characteristics of modal, creaky and breathy vocal characteristics based upon the analysis of glottal flow waveforms obtained by inverse filtering of speech signal. Lee and Childers (1989) have described several vocal disorders using time and frequency glottal flow characteristics based upon the observations of several glottal flow waveforms. A brief review of glottal flow characteristics for various vocal disorders is given in Chapter 4.

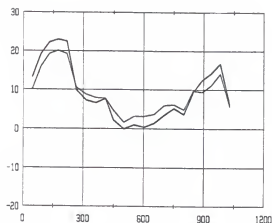
It is difficult to estimate a glottal flow waveform from the speech signal. Inverse filtering techniques using LPC analysis [Davis, 1976; Lee and Childers, 1989] require numerous computations. Other methods, such as Rothenberg's flow mask [Rothenberg, 1973; Holmberg et al., 1988] and Sondhi's reflectionless tube [Sondhi, 1975], require additional equipment for data collection, which may lead to unnatural phonations by human subjects. Also, these methods cannot completely remove the effect of vocal tract resonances from the speech signal, especially from a speech signal recorded from subjects with vocal disorders. Recently, Childers and Ting (1991) have developed an inverse filtering procedure based upon an adaptive signal processing technique that performs better than some conventional normal inverse filtering techniques using pitch-synchronous data.

The procedure to measure HNR described above can also be used to measure the degradation of high-frequency harmonics in glottal flow waveforms. In order to differentiate the two, we define  $HNR_s$  as the Harmonic to Noise measure obtained from the speech signal and  $HNR_g$  as the Harmonic to Noise measure obtained from the glottal flow waveform. The  $HNR_s$  and  $HNR_g$  may not be proportional to each other since, the transfer function of the vocal tract amplifies and attenuates different frequencies in different proportions. If we hypothesize that the degradation of high-frequency harmonics in the speech signal is due to a degradation of high-frequency harmonics in the glottal source, the  $HNR_g$  measure can be used as a

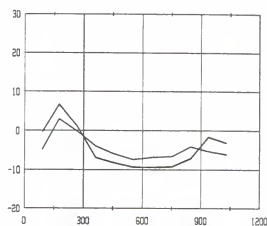
significant glottal factor for modeling and predicting the severity of various vocal disorders.

In Figure E-2 and Figure E-3, the variation of the HNR measured from a segment of a speech signal of sustained vowel /t / and from the corresponding inverse filtered glottal flow waveform of a breathy phonation are shown. Instead of calculating the average value of  $\text{HNR}_g$  and  $\text{HNR}_s$  measures for the entire speech and glottal flow waveform, we have shown the variation in the value of these measures with time. From this type of representation we can study the time varying changes in the HNR measure in the speech signal and glottal flow waveform. The two curves in each figure show the variation of HNR when measured from the harmonics above the fundamental frequency and from the harmonics above 2 KHz. It can be observed that, for this phonation the variation in the HNR measure is not significantly different for a different starting harmonic. From these figures, it can also be observed that the variation of the HNR with time is smoother if eight pitch periods are used instead of four pitch periods. In other words, the smaller the number of pitch periods, the lesser is the averaging of data used for measuring the HNR and the better is the observation of variation of the HNR with time in a phonation. The dynamic range and pattern of variation of HNR, when calculated using the FFT, is almost the same as that calculated using the DFT. It can be observed that the  $\text{HNR}_g$ , although not proportional to the  $\text{HNR}_s$ , shows a similar trend as the  $\text{HNR}_s$ .

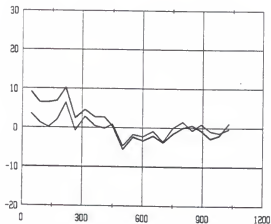
A measure of degradation of high-frequency harmonics should be sensitive to the presence of aspiration noise, pitch period perturbation and amplitude perturbation and not to the fundamental frequency of the speech signal or the glottal flow waveform. Muta et al., (1988) have shown that NSR was less sensitive to variations in fundamental frequency than was the HNR developed by Yumoto et al. (1982). We decided to test the sensitivity of the HNR to variation in fundamental frequency using several model generated glottal source waveforms and synthetic speech segments for which only the



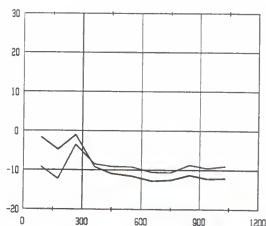
[a]



[c]

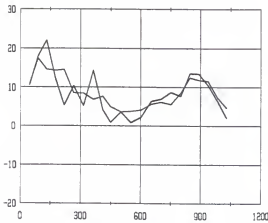


[b]

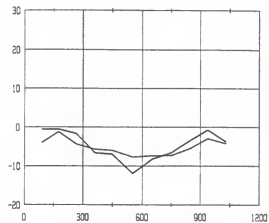


[d]

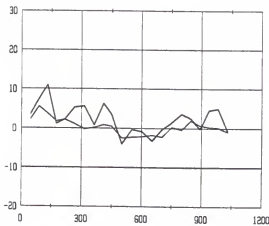
Figure E-2: Variation of  $HNR_s$  in speech signal with time (time is in ms and HNR magnitude is in dB)  
 a) Calculated from four pitch periods using FFT  
 b) Calculated from four pitch periods using DFT  
 c) Calculated from eight pitch periods using FFT  
 d) Calculated from eight pitch periods using DFT



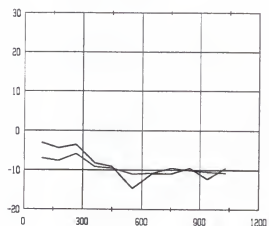
[a]



[c]



[b]



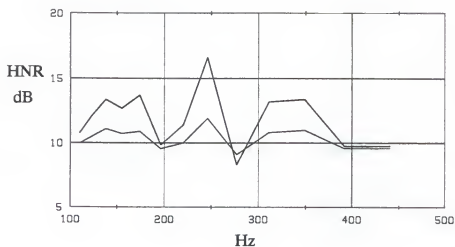
[d]

Figure E-3: Variation of  $HNR_g$  in glottal flow waveform with time (time is in ms and HNR magnitude is in dB)  
 a) Calculated from four pitch periods using FFT  
 b) Calculated from four pitch periods using DFT  
 c) Calculated from eight pitch periods using FFT  
 d) Calculated from eight pitch periods using DFT

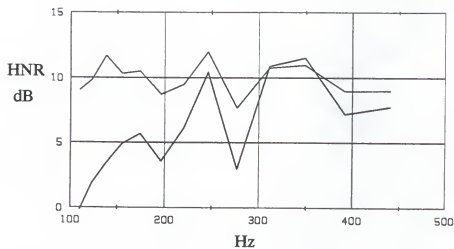


fundamental frequency was varying. The SNR (Signal to Noise Ratio) was kept constant at 30 dB for all the glottal source waveforms and synthetic speech segments. We used the flexible formant synthesizer and the new glottal source model for generating these glottal source waveforms and synthetic speech segments. In Figure E-4 the variation in the value of  $HNR_g$ , when the fundamental frequency was varied from 110 Hz to 440 Hz at six logarithmic steps per octave, is shown. The  $HNR_g$  was calculated from the spectrum obtained by calculating the DFT of four and eight pitch periods for each glottal source waveform. The upper curve in each figure shows the variation of  $HNR_g$  when measured from the harmonics above the fundamental frequency and the lower curve shows the variation of  $HNR_g$  when calculated from the harmonics above 2 KHz. From these figures it can be observed that the larger the number of pitch periods used for calculating  $HNR_g$ , the lesser is the sensitivity to variations in fundamental frequency. Also,  $HNR_g$  is more sensitive to a variation in fundamental frequency when calculated from the harmonics above 2 KHz. However, as mentioned earlier, the appropriate values for the number of pitch periods and the starting harmonic for the calculation of  $HNR_s$  and  $HNR_g$  are those, for which, these measures are most effective in classifying vocal disorders.

In the Figure E-5a the variation of  $HNR_g$  when calculated from the spectrum obtained by using the FFT instead of the DFT is shown. It can be observed that as the fundamental frequency increases, the value of  $HNR_g$  increases. This is due to the fact that as the fundamental frequency increases, the bandwidth of each harmonic increases. The number of samples of spectrum obtained by FFT in each harmonic increases, resulting in an increase in the value of  $HNR_g$ . One method to reduce this artifact is to normalize the  $HNR_g$  (and  $HNR_s$ ) with the fundamental frequency. Accordingly, the definition of HNR is modified as follows:

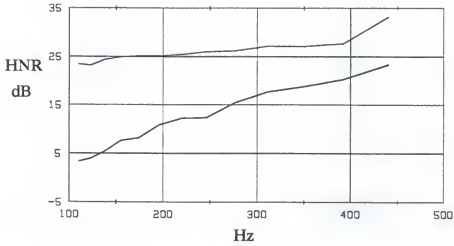


[a]

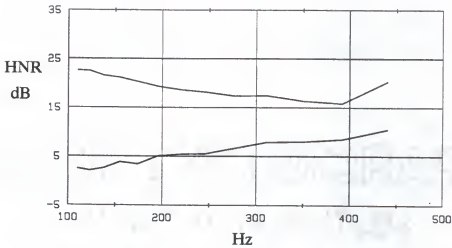


[b]

Figure E-4: Variation in HNR due to variation in  $F_0$   
 (HNR calculated using DFT)  
 a) HNR calculated from 4 glottal source pulses  
 b) HNR calculated from 8 glottal source pulses



[a]



[b]

Figure E-5: Variation in HNR due to variation in F0  
 (HNR calculated from FFT of 8 glottal source pulses )  
 a) Not normalized by the fundamental frequency  
 b) Normalized by the fundamental frequency

$$HNR = 20.0 * \log \left[ \frac{\sum_i H_i}{F0 \sum_i N_i} \right] \quad E-4$$

From Figure E-5b, we observe that the normalization reduces the sensitivity of  $HNR_g$  to variations in fundamental frequency. From a comparison of Figure E-4b and Figure E-5b we observe that the sensitivity of the normalized  $HNR_g$  to the variation of fundamental frequency when calculated using FFT is almost the same as that when  $HNR_g$  is calculated using DFT. The  $HNR_s$ , although not proportional to  $HNR_g$ , showed similar trend of variation as the  $HNR_g$  for each of the cases shown in Figure E-4 and Figure E-5.

The  $HNR_g$  can be used as one of the significant glottal factors for modeling/synthesizing various vocal characteristics. From a modeling/synthesis point of view,  $HNR_g$  can be used as an index for measuring: 1) high-frequency aspiration noise, 2) pitch period perturbation and 3) amplitude perturbation in glottal source pulses in the speech signal. As described in Chapter 4, these are considered to be the main characteristics of breathy, rough and hoarse vocal characteristics. By systematically controlling this measure, we can synthesize speech tokens with varying degrees of breathiness, roughness and hoarseness. The method to systematically vary this glottal factor using the new glottal source model is described in Chapter 5.

## REFERENCES

- Allen, E. L., and Hollien, H. (1973). "A laminagraphic study of pulse (vocal fry) register phonation," *Folia Phoniatrica* 25, 241-250.
- Ananthapadmanabha, T. V. (1984). "Acoustic analysis of voice source dynamics," in *Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden*, Vol. 2-3, 1-24.
- Ananthapadmanabha, T. V., and Fant, G. (1982). "Calculation of true glottal flow and its components," *Speech Comm.* 1, 167-184.
- Atal, B. S., and Hanauer S.L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.* 50(2), 637-655.
- Atal, B.S., and Remde, J. S. (1982). "A new model of LPC excitation for producing natural-sounding speech at low bit rates," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Paris, France, 614-617.
- Boone, D.R. (1971). *The Voice and Voice Therapy*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Bristow, G. (1984). *Electronic Speech Synthesis*, McGraw-Hill Book company, New York.
- Childers, D. G., Hicks, D. M., Moore, G. P., Eskenazi, L., and Lalwani, A. L. (1990) "Electroglottography and vocal fold physiology." *J. Speech and Hearing Res.* 33, 245-254.
- Childers, D. G., and Krishnamurthy, A. K. (1985). "A critical review of Electroglottography," *CRC Critical Reviews in Biomedical Engineering* 12(2), 131-164
- Childers, D. G., Naik, J. M., Larar, J. N., Krishnamurthy, A. K., and Moore, G. P. (1983). "Electroglottography, speech and ultra-high speed cinematography," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by Titze, I. R. and Scherer, R. C., The Denver Center for the Performing Arts, Inc., Denver, Colorado, 202-220.
- Childers, D. G., and Ting, Y. T. (1992). "Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor," submitted.
- Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis-synthesis," *Speech Comm.* 9, 97-117.

- Coleman, R. F. (1960). "Some acoustic correlates of hoarseness," Master's thesis, Vanderbilt University, Nashville, Tennessee.
- Cranen, B., and Boves, L. (1985). "A set-up for testing the validity of the two mass model of the vocal folds," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by Titze, I. R. and Scherer, R. C., The Denver Center for the Performing Arts, Inc., Denver, Colorado, 500-513.
- d'Alessandro, C. (1990). "Time-frequency speech transformation based on an elementary waveform representation," *Speech Comm.* 9, 419-431.
- Davis, B. S. (1976). "Computer evaluation of laryngeal pathology based on inverse filtering of speech," *Speech Communication Research Laboratory, Monograph Number 13*, Santa Barbara, CA.
- Dunn, H. K., and White, S. D. (1940). "Statistical measurements of conversational speech," *J. Acoust. Soc. Am.* 23, 278-288.
- Eskenazi, L. (1988). "Acoustic correlates of voice quality and distortion measures for speech processing," Ph.D. dissertation, University of Florida, Gainesville.
- Eskenazi, L., Chiders, D. G., and Hicks, D. M. (1990). "Acoustic correlates of vocal quality," *J. Speech and Hearing Res.* 33, 298-306.
- Fant, G. (1956). "On predictability of formant levels and spectrum envelopes from formant frequencies," in *Speech Synthesis*, edited by Flanagan, J. L. and Rabiner, L. R., Dowden, Hutchinson and Ross, Inc., Stroudsburg, Pennsylvania, 216-228.
- Fant, G. (1960). *Acoustic Theory of Speech Production*, Mouton, Paris.
- Fant, G., and Ananthapadmanabha, T.V. (1982). "Truncation and superposition," in *Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden* 2-3, 1-17.
- Fant, G., Ishizaka, K., Lindqvist, J., and Sundberg, J. (1972). "Glottal source and excitation analysis," *Quarterly Progress and Status Report, Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden*, 4, 1-13.
- Fant, G., and Lin, Q-G. (1988). "Frequency domain interpretation and derivation of glottal flow parameters," *Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden*, 2-3, 1-21.
- Fant, G., Liljencrants, J., and Lin, Q-G. (1985). "A four parameter model of glottal flow," paper presented at the French-Swedish Sympo., Grenoble, France.
- Flanagan, J. L. (1957). "Note on the design of terminal-analog speech synthesizers," *J. Acoust. Soc. Am.* 29, 306-310.
- Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception*, 2nd ed., Springer-Verlag, New York.

- Flanagan, J. L., and Golden, R. M. (1966). "Phase vocoder," *Bell System Techn. J.* 65(5), 747-770.
- Flanagan, J. L., Coker, C. H., and Bird, C. M. (1956). "Computer simulation of a formant-vocoder synthesizer," in *Speech Synthesis*, edited by Flanagan, J. L. and Rabiner, L. R., Dowden, Hutchinson and Ross, Inc., Stroudsburg, Pennsylvania, 244-254.
- Flanagan, J. L., and Landgraf, I. L. (1968). "Self-oscillating source for vocal tract synthesizers," *IEEE Trans. Audio and Electroacoustics* 16, 57-64.
- French, N. R., and Steinberg, J. C. (1947). "Factors governing intelligibility of speech sounds," *J. Acoust. Soc. Amer.* 19, 90-119.
- Fujisaki, H., and Ljungqvist, M. (1986). "Proposal and evaluation of models for the glottal source waveform," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Tokyo, Japan, 1605-1608.
- Gobl, C. (1988). "Voice source dynamics in connected speech," in *Quarterly Progress and Status Report*, Royal Institute of Technology, Stockholm, Sweden, 1, 123-159.
- Gobl, C. (1989). "A preliminary study of acoustic voice quality correlates," in *Quarterly Progress and Status Report*, Royal Institute of Technology, Stockholm, Sweden, 4, 9-22.
- Gold, B., and Rabiner, L. R. (1968). "Analysis of digital and analog formant synthesizers," *IEEE Trans. Acoust., Speech, and Signal Processing* AU-16, No. 1 81-94.
- Hirano, M. (1981). *Clinical Examination of Voice*, Springer-Verlag, New York.
- Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S., and Tanabe, M. (1984). "Harmonic-intensity analysis of normal and hoarse voices," *J. Acoust. Soc. Am.*
- Hamlet, S. L., (1981). "Ultrasound assessment of phonatory function," *Conf. on Assessment of Vocal Pathology*, ASHA Reports, 11, 128-140.
- Hecker, M. H., and Kreul, E. J. (1971). "Descriptions of speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency," *J. Acoust. Soc. Amer.* 49, 1275-1282.
- Hollien, H. (1974). "On vocal register," *Journal of Phonetics* 2, 125-144.
- Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *J. Acoust. Soc. Am.* 84(2), 511-529.
- Holmes, J. N. (1962). "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," *Proc. IV Internat. Congress on Acoustics*, Copenhagen, Denmark, 1-4

- Holmes, J. N. (1973). "The Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoustics* 21, 298-305.
- Holmes, J. N. (1983). "Formant synthesizers: cascade or parallel," *Speech Comm.* 2(4), 251-274.
- Holmes, W. J., Holmes, J. N., and Judd, M. W. (1990). "Extension of the bandwidth of the JSRU parallel formant synthesizer for high quality synthesis of male and female speech," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc., Albuquerque, N.M.*, 313-316.
- Horii, Y. (1980). "Vocal shimmer in sustained phonation," *J. Speech and Hearing Res.* 23, 202-209.
- Ishizaka, K., and Flanagan, J. L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Techn. J.* 51(6), 1233-1268.
- Karlsson, I. (1986). "Glottal waveforms for normal female speakers," *Journal of Phonetics* 14, 415-419.
- Kernigham, B. W., and Ritchie, D.M. (1978). *The C Programming Language*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632.
- Kitzing, P. (1982). "Photo- and electrophysiological recording of the laryngeal vibratory pattern during different registers," *Folia Phoniatrica* 34, 234-241.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* 67(3), 971-995.
- Klatt, D. H. (1986). "Detailed spectral analysis of a female voice," *J. Acoust. Soc. Am. Suppl.* 80(1), S97.
- Klatt, D. H. (1987). "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.* 82(3), 737-793.
- Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.* 87(2), 820-857.
- Kobayashi, T., and Sekine, H., (1990) "Statistical properties of fluctuation of pitch intervals and its modeling for natural synthetic speech," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing, Albuquerque, N.M.*, 321-324.
- Koike, Y., Takahashi, H., and Calcaterra, T. C. (1977). "Acoustic measures for detecting laryngeal pathology," *Acta Otolaryngologica* 84, 105-117.
- Kojima, H., Gould, W. J., and Lambiase, A., and Isshiki, N. (1980). "Computer analysis of hoarseness," *Acta Otolaryngologica* 89, 547-554.
- Lalwani, A. L. (1991). "The LF model," *International Report, Mind-Machine Interaction Research Center, the University of Florida, Gainesville.*



- Lalwani, A. L., and Childers, D. G. (1991a). "Modeling vocal disorders via formant synthesis," Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing, Toronto, Canada, 505-508.
- Lalwani, A. L., and Childers, D. G. (1991b). "A flexible formant synthesizer," Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing, Toronto, Canada, 777-780.
- Laver, J., and Hanson, R. (1981). "Describing the normal voice," in *Evaluation of Speech in Psychiatry*, edited by J. Darby, Grune and Stratton, Inc., New York, pp. 51-78.
- Lawrence, W. (1953). "The synthesis of speech from signals which have a low information rate," in *Speech Synthesis*, edited by Flanagan, J. L. and Rabiner, L. R., Dowden, Hutchinson and Ross, Inc., Stroudsburg, Pennsylvania, 234-243.
- Lee, C. K. (1988). "Voice quality: analysis and synthesis," Ph.D. Dissertation, University of Florida, Gainesville.
- Lee, C. K., and Childers, D. G. (1989). "Some acoustical perceptual and physiological aspects of vocal quality," paper presented at Vocal Fold Physiology Conference, Stockholm, Sweden.
- Lieberman, P. (1961). "Perturbation in vocal pitch," J. Acoust. Soc. Amer. 33, 597-603.
- Lieberman, P. (1963). "Some acoustic measures of the fundamental periodicity of normal and pathological larynges," J. Acoust. Soc. Amer. 35, 344-353.
- Maeda, S. (1982). "A Digital Simulation Method of the Vocal-Tract System," Speech Comm. 1(3), 199-229.
- Malah, D. (1979). "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signal," IEEE Trans. Acoust., Speech, and Signal Processing 27(2), 121-133.
- Markel, J. D., and Gray, A. H. (1976). *Linear Prediction of Speech*, Springer-Verlag, Berlin.
- Monsen, R. B., and Engebretson, A. M. (1977). "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am. 62(4), 981-993.
- Moore, G. P. (1975). "Observation on the physiology of hoarseness," Proc. 4th Int. Congress of Phonetic Sci., Helsinki, Finland, 92-95.
- Moulines, E., and Charpenter, F. (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Comm. 9, 453-467.
- Muta, H., T. Baer, Wagatsuma, K., Muraoka, T., Fukuda, H. (1988). "A pitch-synchronous analysis of hoarseness in running speech," J. Acoust. Soc. Amer. 84(4), 1292-1301.

- Oppenheim, A. V., and Schafer, R. W. (1978). *Digital Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in the study of vowels," *J. Acoust. Soc. Am.* 24, 175-184.
- Pinto, N. B. (1987). "A high-fidelity speech synthesizer," Master's thesis, University of Florida, Gainesville.
- Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant speech synthesis: improving production quality," *IEEE Trans. Acoust., Speech, and Signal Processing* 37(12), 1870-1887.
- Pinto, N. B., and Titze, I. R. (1990). "Unification of perturbation measures in speech signals," *J. Acoust. Soc. Am.* 87(3), 1278-1289.
- Rabiner, L. R. (1968). "Digital-formant synthesizer for speech synthesis studies," *J. Acoust. Soc. Am.* 43, 822-828.
- Rabiner, L. R., Jackson, L. B., Schafer, R. W. and Coker, C. H. (1971). "A hardware realization of a digital formant synthesizer," in *Speech Synthesis*, edited by Flanagan, J. L., and Rabiner, L. R., Dowden, Hutchinson and Ross, Inc., Stroudsburg, Pennsylvania, 262-266.
- Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- Rothausen, E.H., Urbauer, G.E., and Pacht, W.P. (1971). "A comparison of preference measurement methods," *J. Acoust. Soc. Amer.* 49, 1291-1308.
- Rothenberg, M. R. (1973). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.* 53, 1632-1645.
- Rothenberg, M. R. (1985). "Source-tract acoustic interaction in breathy voice," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by I. R. Titze and R. C. Scherer, The Denver Center for the Performing Arts, Inc., Denver, Colorado, pp. 465-481.
- Schroeder, M. R., and Atal, B. S. (1985). "stochastic coding of speech signals at very low bit rates: the importance of speech perception," *Speech Comm.* 4(1-3), 155-162.
- Seneff, S. S. (1982). "Speech transformation without pitch extraction," *IEEE Trans. Acoust., Speech, and Signal Processing* 30(4), 566-578.
- Sondhi, M. M. (1975). "Measurement of the glottal waveform," *J. Acoust. Soc. Amer.* 57(1), 228-232.
- Stella, M. (1985). "Speech synthesis," in *Computer Speech Processing*, edited by Fallside, F., and Woods, W. A., Prentice-Hall International Ltd., London, 421-460.

- Stevens, K. H. (1971). "Airflow and turbulence noise for fricative and stop consonants: static considerations," *J. Acoust. Soc. Am.* 50(4), 1180-1192.
- Titze, I. R. (1984). "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Am.* 75(2), 520-580.
- van den Berg, J. W. (1968). "Mechanism of the larynx and the laryngeal vibrations," in *Fonts of Phonetics*, edited by J. Malmberg, North-Holland, London, pp. 278-308.
- Verhelst, W., and Nilens, P. (1986). "A modified-superposition speech synthesizer and its applications," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Tokyo, Japan, 2007-2010.
- Wendahl, R. W. (1963). "Laryngeal analog synthesis of harsh voice quality," *Folia Phoniatica* 15, 241-250.
- Wendahl, R. W. (1966). "Laryngeal analog synthesis of jitter and shimmer\_auditory parameters of harshness," *Folia Phoniatica* 18, 98-108.
- Wolf, V. I., and Steinfatt, T. M. (1987). "Prediction of vocal severity within and across voice types," *J. Speech and Hearing Res.* 30, 230-240.
- Wong, C-F. (1991). "The incorporation of glottal source-vocal tract interaction effects to improve the naturalness of synthetic speech," Ph.D. Dissertation, University of Florida, Gainesville.
- Wong, D. Y., and Markel, J. D. (1978). "An excitation function for LPC synthesis which retains the human glottal phase characteristics," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Tulsa, Oklahoma, pp. 171-174.
- Yanagihara, H. (1967). "Significance of harmonic changes and noise components in hoarseness," *J. Speech and Hearing Res.* 10, 531-541.
- Yumoto, E., Gould, W., and Baer, T. (1982). "Harmonic-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.* 71, 1544-1550.
- Yumoto, E., Sasaki, Y., Okamura, H. (1984). "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," *J. Speech and Hearing Res.* 27, 2-6.

## BIOGRAPHICAL SKETCH

Ajit L. Lalwani was born in Poona, India, on May 26, 1962. He received the Bachelor of Engineering in electrical engineering from the College of Engineering, Poona in 1983. He worked for a year as a Trainee Engineer in Buckau Wolf India Ltd, Poona.

Then he joined the University of Florida in Gainesville, Florida, for his graduate study, where his principal area of study is digital signal processing. He received his Master of Engineering in December, 1986. For a year he worked as a research assistant in the Electrophysiology Lab, where his research interest was processing evoked responses in EEG signal. He has been a member of Mind-Machine Interaction Research Center since 1984, where his research interest has been in speech signal processing, mainly focused on speech analysis/synthesis and modeling.

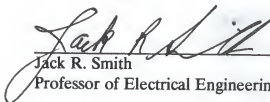
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



---

Donald G. Childers, Chairman  
Professor of Electrical Engineering

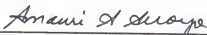
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



---

Jack R. Smith  
Professor of Electrical Engineering

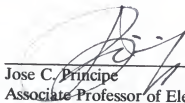
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



---

Amauri A. Arroyo  
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



---

Jose C. Principe  
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



---

Howard B. Rothman  
Professor of Communication Processes and  
Disorders

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 1992



---

Winfred M. Phillips  
Dean, College of Engineering

---

Madelyn M. Lockhart  
Dean, Graduate School